Université de Montréal

L’évolution des pangénomes de procaryotes sur des échelles de temps humaines

Par

Arnaud N’Guessan

Département de biochimie et médecine moléculaire

Faculté de médecine

Mémoire présenté en vue de l’obtention du grade de

Maître ès sciences (M.Sc.)

en bio-informatique

3 décembre 2020

Université de Montréal

Département de biochimie et médecine moléculaire, Faculté de médecine

---

Ce mémoire intitulé

L'évolution des pangénomes de procaryotes sur des échelles de temps humaines

Présenté par

Arnaud N'Guessan

A été évalué(e) par un jury composé des personnes suivantes

Franz Lang
Président-rapporteur

Adrian Serohijos
Directeur de recherche

Jesse Shapiro
Codirecteur

François-Joseph Lapointe
Membre du jury

Résumé

Le pangénome est l'ensemble des gènes uniques retrouvé chez une espèce. Dans le cas des espèces procaryotes, notamment celles qui sont présentes dans le microbiote intestinal humain, la variation du contenu en gène est caractérisée par des événements de gain de gènes principalement par transfert horizontal de gènes (THG) et de perte de gène. Cette variation du contenu en gène peut être plus rapide que le taux de mutation et permettre aux microbes de s'adapter rapidement à des pressions sélectives. Cela justifie donc l'étude de l'évolution pangénomique des procaryotes sur des échelles de temps humaines qui sont considérées comme étant courtes du point de vue évolutif, par exemple de l'ordre de quelques années. La plupart des études sur ce sujet impliquent des espèces relativement distantes qui ont divergé depuis des millions d'années. De plus, l'équilibre des forces évolutives majeures impliquées, telles que le THG, la sélection, la dérive génétique et les mutations, n'est pas clairement défini et est au cœur d'un débat dans la littérature. Ce projet de maîtrise permet donc d'élargir le portrait évolutif des pangénomes de procaryotes en s'intéressant à l'évolution des gènes transférés horizontalement, aussi appelés gènes mobiles, sur de courtes échelles de temps. Pour ce faire, nous allons d'abord passer en revue la littérature pertinente en lien avec ce sujet, notamment les méthodes employées pour détecter les gènes mobiles et les modèles d'évolution pangénomique. Nous allons ensuite analyser l'évolution d'une collection de 37 853 gènes mobiles impliqués dans des THG récents détectés dans le microbiote intestinal d'individus provenant d'Amérique du Nord ou des îles Fidji. Pour détecter des signatures évolutives des forces en action, nous estimerons divers paramètres de génétique des populations à partir de l'alignement entre les lectures de séquençage métagénomique de 176 microbiotes fidjiens et cette collection de gènes mobiles. Nous expliquerons aussi l'outil de simulations évolutives que nous avons développé afin de valider et expliquer certaines de nos observations. Sans exclure la présence de pressions de sélection pour des gènes mobiles ayant des fonctions spécifiques, les données réelles et les simulations nous amènent à conclure que l'évolution des gènes mobiles sur de courtes échelles de temps peut être expliquée par un modèle d'évolution où les gènes mobiles ne sont pas largement adaptifs à leurs hôtes humains ou microbiens, contrairement à ce qui est parfois observé sur de longues échelles de temps évolutif. **Mots-clés** : Pangénome, évolution, gènes mobiles, transfert horizontal, microbiote intestinal humain, procaryotes, simulation.

Abstract

The pangenome is the collection of unique genes found in a species. For prokaryotes, especially those present in the human gut microbiota, variation in gene content is characterized by gene gain through horizontal gene transfer (HGT) and gene loss. In human gut, gene content variations can occur at faster rates than mutation, which allow microbes to adapt rapidly to environmental changes. This justifies the study of the prokaryotes pangenome evolution on human time scales which are considered evolutionarily short, e.g. in the order of few years. Most studies about the evolution of prokaryotic pangenomes involve relatively distant species that have diverged since millions of years. In addition, the balance of major evolutionary forces involved, such as horizontal transfer, selection, genetic drift, and mutations, is not clearly defined and is debated in literature. This master's project therefore aims to broaden the evolutionary portrait of prokaryotic pangenome evolution by focusing on near-term evolution. To do this, we will first review the relevant literature related to this topic, including the methods used to detect mobile genes and the pangenome evolution models. We will then analyze the evolution of a pre-existing collection of 37 853 mobile genes involved in recent HGT events detected in the gut microbiota of individuals from North America and Fiji Islands. To detect evolutionary signatures of the forces in action, we will estimate various population genetics parameters from the alignment between metagenomic sequencing reads of 176 Fijian microbiomes and this collection of mobile genes. We will also explain the evolutionary simulation tool that we have developed in order to validate and explain some of our observations. While we don’t exclude the importance of selection for specific cellular functions for pangenome evolution, we found that the near-term evolution of mobile genes can be explained by a model in which mobile genes can spread selfishly without being largely adaptive to their human or microbial hosts, contrarily to what is often observed over longer evolutionary time scales.
Keywords: Pangenome, Evolution, Mobile genes, Horizontal Gene Transfer, Human gut microbiome, Prokaryotes and Simulation.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ADN: Acide désoxyribonucléique

AMP: Adénosine monophosphate

C: Cytosine

CAI: *Codon adaptation index*

COG: *Cluster of orthologous genes*

dbCAN: *Carbohydrate-active enzyme annotation database*

FijiCOMP: *Fiji community microbiome project*

FRQNT: Fonds de recherche du Québec – Nature et technologies

G: Guanine

HGT: *Horizontal gene transfer*

HMP: *Human microbiome project*

KEGG: *Kyoto encyclopedia of genes and genomes*

KS: Kolmogorov-Smirnov

LRT: *Likelihood ratio test*

NSERC: *Natural sciences and engineering research council of Canada*

PFAM: *Protein family database*

SNV: *Single nucleotide variant*

SPR: *Subtree pruning and regrafting*

THG: Transfert horizontal de gènes

TIGRFAM: *The institute for genomic research protein family database*

*Les mots en anglais sont représentés en italique

Je dédie ce mémoire à mes parents qui ont tant donné afin que je puisse poursuivre mes rêves.

Remerciements

Tout d'abord, je tiens à remercier mon directeur de recherche, Dr. Adrian Serohijos, et mon co-directeur de recherche, Dr. Jesse Shapiro, pour leur soutien, leurs conseils, l'opportunité de travailler sur un projet aussi stimulant et surtout la confiance qu'ils ont en moi. En plus d'être 2 chercheurs brillants, ce sont 2 personnes attachantes et exemplaires qui me servent de modèles et avec qui il est facile de travailler et apprendre.

Je tiens aussi à remercier les membres de mes 2 laboratoires ainsi que ceux du laboratoire du Dr. Stephen Michnick pour les commentaires constructifs et les présentations de qualité qui ont contribué à mon apprentissage. Je remercie particulièrement Pouria Dasmeh, Sébastien Boyer et Louis Gauthier pour les discussions et le partage de leurs connaissances.

Je remercie aussi les membres du jury pour le temps accordé à l'évaluation de mon mémoire et les commentaires constructifs qui m'aideront à progresser. Je suis aussi reconnaissant envers l'Université de Montréal, Calcul Canada, le Conseil de recherches en sciences naturelles et en génie du Canada (NSERC) et les fonds québécois de la recherche sur la nature et les technologies (FRQNT) pour les formations offertes, les ressources financières et les ressources matérielles qui m'ont été allouées afin de mener à bien mon projet de maîtrise.

Finalement, je remercie ma famille et mes amis pour leur soutien et les bons moments passés ensemble, plus particulièrement mes parents et mon ami Adi.

Avant-propos

L'évolution des pangénomes de procaryotes est actuellement une source de débat dans la littérature. Au centre de ce débat, se trouve l'opposition entre un modèle adaptatif d'évolution, c'est-à-dire un modèle où la sélection a plus d'effets que la dérive génétique sur le contenu en gène des espèces, et un modèle neutre ou non adaptatif d'évolution. Non seulement il n'y a pas encore de consensus quant à ce sujet, mais en plus, l'évolution des pangénomes de procaryotes a majoritairement été étudiée sur de longues échelles de temps évolutif, par exemple entre des espèces ayant divergé il y a des millions d'années (Andreani, Hesse, & Vos, 2017; Bobay & Ochman, 2018; Brito et al., 2016; McInerney, McNally, & O'Connell, 2017; Sela, Wolf, & Koonin, 2016). Cependant, des échelles de temps évolutif plus courtes sont tout aussi pertinentes puisque les microbes peuvent s'adapter rapidement et développer des phénotypes comme la résistance aux antibiotiques dans des échelles de temps humaines, c'est-à-dire de l'ordre de quelques mois ou années (Garud & Pollard, 2020; Jiang, Hall, Xavier, & Alm, 2019). De plus, la plupart des études mentionnées précédemment analysent la variation du contenu en gène et le THG à l'aide de paramètres mesurés à l'échelle de génomes entiers. Cependant, différents gènes mobiles n'évoluent pas nécessairement de la même façon (Koonin & Wolf, 2010; Shapiro, 2017).

Ainsi, nous avons décidé d'étudier l'évolution des pangénomes de procaryotes sur de courtes échelles de temps évolutif tout en tenant compte des variations dans l'évolution des différents gènes mobiles. Plus précisément, nous nous sommes intéressés à une collection de 37 853 gènes mobiles impliqués dans des THG récents détectés dans le microbiote intestinal d'individus provenant d'Amérique du Nord ou des îles Fidji. Pour détecter des signatures des forces évolutives en action, nous avons estimé divers paramètres de génétique des populations à partir de l'alignement entre les lectures de séquençage métagénomique de 176 microbiotes fidjiens et cette collection de gènes mobiles. Nous avons ensuite intégré un module de simulation de THG et de perte de gènes à un simulateur d'évolution microbienne afin de valider nos conclusions.

Le présent document a pour objectif de présenter ces résultats de recherche, la méthodologie employée, le sujet de recherche, l'outil de simulation évolutive développé, une discussion sur les éléments précédents et une perspective sur les travaux futurs.

1 Revue de littérature

1.1 Introduction

Le pangénome reflète la diversité du contenu en gène d'une espèce. Dans le cas des microbes, son évolution dépend non seulement de facteurs écologiques, comme l'abondance des espèces dans une niche écologique, mais aussi de l'équilibre des forces évolutives comme la sélection, la dérive génétique, les mutations, la recombinaison génétique, la perte de gène et le THG (Garud & Pollard, 2020). Un milieu dans lequel cet équilibre n'est pas encore clairement résolu, notamment sur de courtes échelles de temps, et qui démontre clairement l'importance de l'évolution pangénomique microbienne est le microbiote intestinal humain.

Les fonctions et la composition des communautés microbiennes intestinales humaines ont été largement étudiées à travers différentes populations humaines (Brito et al., 2016; Valdes, Walter, Segal, & Spector, 2018). La recherche a souligné l'importance de leurs phénotypes pour la santé humaine. Par exemple, le microbiote intestinal humain contribue à la fermentation des fibres non digestibles, à l'amélioration du métabolisme des lipides et des glucides, à la régulation de l'inflammation intestinale ou même à la production de composés antimicrobiens (Valdes et al., 2018). Les microbes de l'intestin ont une dynamique complexe et échangent relativement fréquemment du matériel génétique par transfert horizontal de gènes (THG). Cela contribue à l'adaptation à leur environnement, notamment par la propagation de la résistance aux antimicrobiens (Jiang et al., 2019), et module l'évolution de leurs pangénomes (McInerney et al., 2017; Sela et al., 2016).

Puisque l'évolution pangénomique a des conséquences sur des phénotypes microbiens d'intérêt clinique et que ce phénomène n'est pas encore clairement compris, il est important de le modéliser et déterminer les facteurs qui ont le plus d'influence sur celui-ci. Le présent chapitre a pour but de définir les concepts essentiels en lien avec l'évolution pangénomique microbienne et expliquer les méthodes et modèles d'analyse existants.

1.2 Méthodes de détection de gènes mobiles

La plupart des études sur l'évolution des pangénomes analysent la variation du contenu en gène et le THG à l'aide de paramètres mesurés à l'échelle de génomes entiers (Andreani et al., 2017; Bobay & Ochman, 2018; McInerney et al., 2017; Sela et al., 2016). Cependant, les différents gènes mobiles n'évoluent pas nécessairement de la même façon (Koonin & Wolf, 2010; Shapiro, 2017). Afin de tenir compte de ces différences ainsi que de leur influence sur l'évolution pangénomique, il faut plutôt effectuer des analyses spécifiques à chaque gène mobile. Pour ce faire, il faut tout d'abord détecter des événements de THG. Les approches permettant d'effectuer cette tâche sont divisées en plusieurs catégories.

1.2.1 Les méthodes paramétriques

La première catégorie de méthode d'inférence de transfert horizontal de gènes est composée de méthodes paramétriques. Les méthodes paramétriques utilisent le concept de signature génomique, c'est-à-dire un paramètre caractérisant la composition ou la structure de séquences d'ADN qui sont typiques à une espèce ou un clade. Plus précisément, l'approche des méthodes paramétriques consiste à trouver des régions génomiques qui ont des signatures génomiques atypiques à l'espèce ou le groupe étudié (**Figure 1.1 p.27**). L'une des signatures génomiques les plus simples à utiliser est la teneur en GC qui est le pourcentage de sites nucléotidiques uniques dans la région génomique contenant les bases G ou C. La teneur en GC est une signature génomique acceptable, car la variabilité du contenu en GC est significative pour les espèces relativement éloignées.

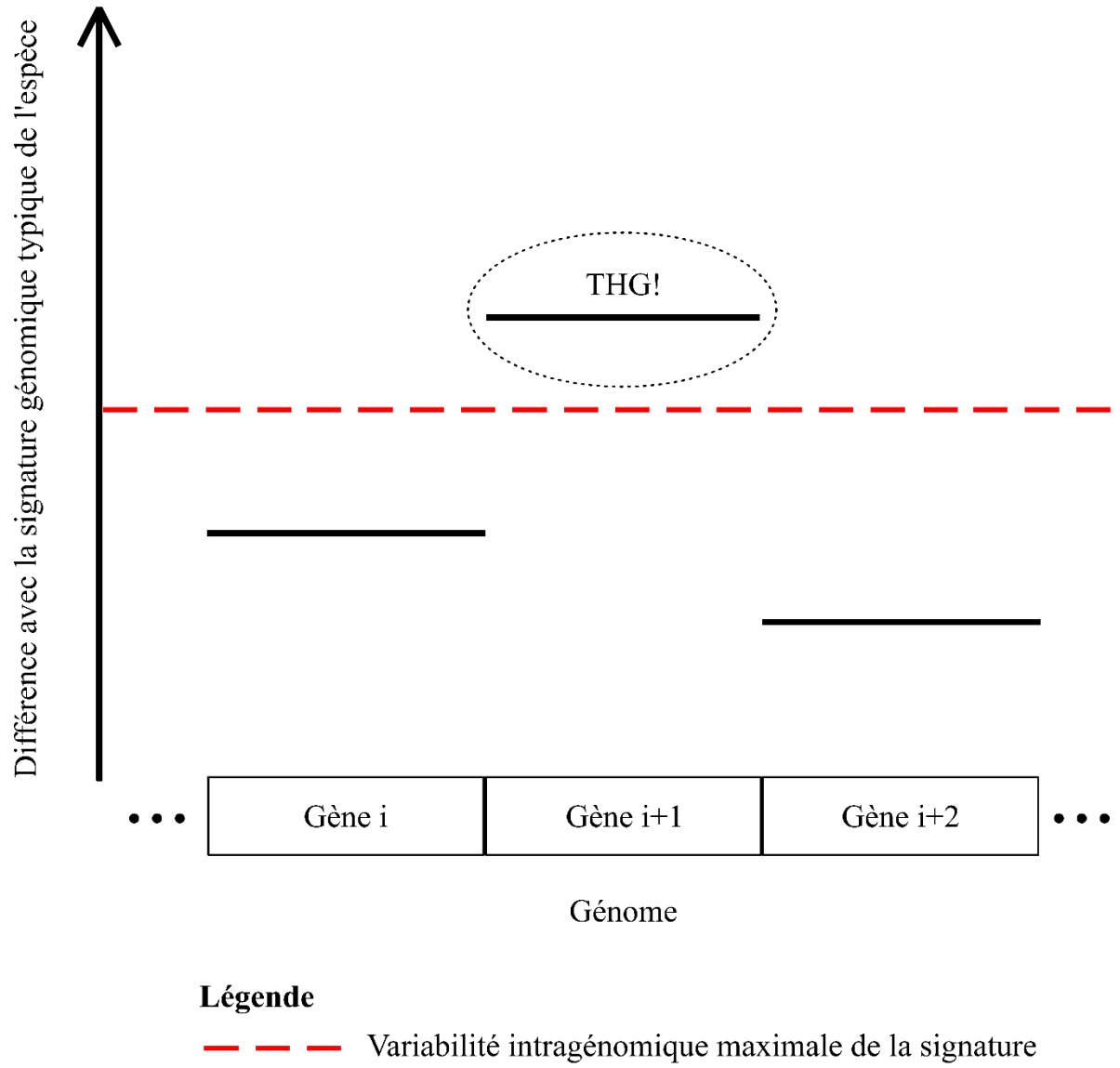


Figure 1.1 Détection d'un gène mobile grâce à une méthode paramétrique

Dans le génome de l'espèce représentée, le gène i+1 a probablement été transféré horizontalement puisque la valeur de sa signature génomique varie significativement par rapport à la signature génomique typique de l'espèce. Les variations significatives de signature génomique permettant d'inférer la présence d'un gène mobile sont celles qui dépassent le seuil de variabilité intragénomique de l'espèce (ligne pointillée rouge).

Cependant, l'utilisation du contenu GC comme signature génomique a ses limites. En effet, la teneur en GC n'est pas nécessairement uniforme à travers le génome d'une espèce en particulier (Ravenhall, Skunca, Lassalle, & Dessimoz, 2015). Par exemple, la teneur en GC de *Tetrahymena thermophila* a tendance à être plus élevée pour les gènes hautement exprimés (Wuitschick & Karrer, 1999). Par conséquent, ne pas être en mesure de différencier le THG de la variabilité intragénomique de la teneur en GC peut augmenter le nombre de faux positifs de cette méthode d'inférence. De plus, les transferts horizontaux entre des espèces étroitement apparentées sont plus difficiles à identifier, car ces espèces ont une signature génomique très similaire, ce qui explique en partie le nombre de faux négatifs. Le phénomène de l'amélioration contribue aussi à expliquer les faux négatifs de cette méthode, puisqu'il se définit par le fait que la signature génomique d'un gène mobile récemment transféré peut être modifiée par des mutations de génération en génération pour finalement s'apparenter à celle de l'espèce réceptrice du gène (Lawrence & Ochman, 1997; Ravenhall et al., 2015).

Les méthodes paramétriques peuvent utiliser d'autres signatures génomiques qui offrent plus de résolution que la teneur en GC même si les inconvénients peuvent être les mêmes, mais avec un biais inférieur. C'est le cas lorsqu'ils sont basés sur la fréquence des oligonucléotides (k-mers), les caractéristiques structurales ou le contexte génomique c'est-à-dire les gènes voisins au gène transféré. Lorsque ces méthodes utilisent la fréquence des oligonucléotides comme signature génomique, les gènes candidats sont scannés en utilisant des fenêtres glissantes et en calculant la fréquence moyenne des différents k-mers. Le calcul de la fréquence moyenne d'utilisation des codons, c'est-à-dire la fréquence à laquelle l'espèce utilise chacun des codons synonymes pour coder chaque acide aminé, et la fréquence des tétranucléotides offrent un bon compromis entre temps de calcul et qualité d'inférence de THG (Dufraigne, Fertil, Lespinats, Giron, & Deschavanne, 2005; Ravenhall et al., 2015).

Quant aux méthodes paramétriques utilisant le contexte génomique, elles détectent des gènes qui sont entourés de séquences qui ne se trouvent généralement pas autour d'eux dans les espèces étudiées. Parmi ces gènes candidats, ceux qui sont également entourés de marqueurs d'éléments mobiles comme des gènes de transposase ou d'intégrase sont encore plus susceptibles d'avoir été impliqués dans un événement de THG (Ravenhall et al., 2015). En effet, les gènes de transposase et d'intégrase codent pour des protéines qui aident à insérer de l'ADN étranger à

plusieurs endroits possibles dans le génome. Un exemple d'application est l'identification d'îlots génomiques c'est-à-dire des éléments génétiques mobiles contenant 10 à 500 kbp qui sont souvent impliqués dans les événements THG et dont la signature génomique est normalement différente de celle des cellules qui les reçoivent. Les algorithmes d'apprentissage automatique peuvent aider à détecter ces îlots génomiques en utilisant le contexte génomique comme signature génomique (Langille, Hsiao, & Brinkman, 2010).

Enfin, il existe des méthodes paramétriques utilisant les caractéristiques structurales de l'ADN génomique, telles que les énergies d'interaction entre des paires de bases voisines ou les angles de torsion comme signature génomique. Plus précisément, ces méthodes recherchent des régions pour lesquelles il existe des différences atypiques avec les patrons de structures périodiques uniques à l'espèce (Worning, Jensen, Nelson, Brunak, & Ussery, 2000).

Malgré les limites des méthodes paramétriques mentionnées précédemment, cette famille de méthodes présente l'avantage d'avoir un coût de calcul inférieur à la plupart des méthodes phylogénétiques et ne nécessite pas nécessairement une séquence de référence de l'espèce étudiée ou des gènes orthologues (Ravenhall et al., 2015). Les méthodes paramétriques ont été les premières à être utilisées pour l'inférence de THG, mais l'amélioration des technologies de séquençage et le développement de la biologie évolutive et des algorithmes de génomique comparative ont permis la création de méthodes phylogénétiques pour l'identification des événements de transfert horizontal (Ravenhall et al., 2015).

1.2.2 Les méthodes phylogénétiques

Les méthodes d'inférence dites phylogénétiques détectent la présence de gènes mobiles grâce aux incohérences entre l'histoire évolutive de ces gènes et celle des espèces qui les contiennent. Dans cette catégorie de méthodes phylogénétiques, il existe deux types d'approches, soit les méthodes phylogénétiques explicites et les méthodes phylogénétiques implicites.

1.2.2.1 Les méthodes phylogénétiques explicites

Les méthodes phylogénétiques explicites peuvent détecter la présence d'un gène mobile en comparant les topologies de l'arbre du gène et à celui des espèces. L'arbre du gène contient les séquences homologues du gène qui sont présentes dans le génome de différentes espèces tandis que l'arbre des espèces est souvent construit à partir de séquences conservées provenant d'une ou plusieurs régions du génome et représente l'histoire évolutive de ces espèces. Les incohérences entre les topologies de ces deux arbres peuvent être expliquées par le THG (**Figure 1.2 p.31**). En revanche, d'autres événements évolutifs peuvent expliquer ces incohérences. Ainsi, les méthodes phylogénétiques explicites considèrent plusieurs scénarios évolutifs et choisissent celui qui optimise un critère de parcimonie ou un modèle probabiliste (Ravenhall et al., 2015).

Les méthodes d'élagage-greffage de sous-arbres, aussi appelé méthodes SPR, et les méthodes de réconciliation sont les méthodes phylogénétiques explicites donnant les meilleures inférences tant en qualité qu'en quantité de détails (Ravenhall et al., 2015). Les méthodes SPR débutent tout d'abord par la suppression des branches de l'arbre du gène qui sont faiblement supportées par l'arbre d'espèce. Après, à partir des sous-arbres générés, il est possible de reconstruire un arbre qui a une topologie similaire à celle de l'arbre des espèces. Il est alors possible de déterminer si le THG concorde avec les changements topologiques apportés. Afin d'augmenter le niveau de confiance des inférences, plusieurs combinaisons d'élagage-greffage sont considérées (Ravenhall et al., 2015).

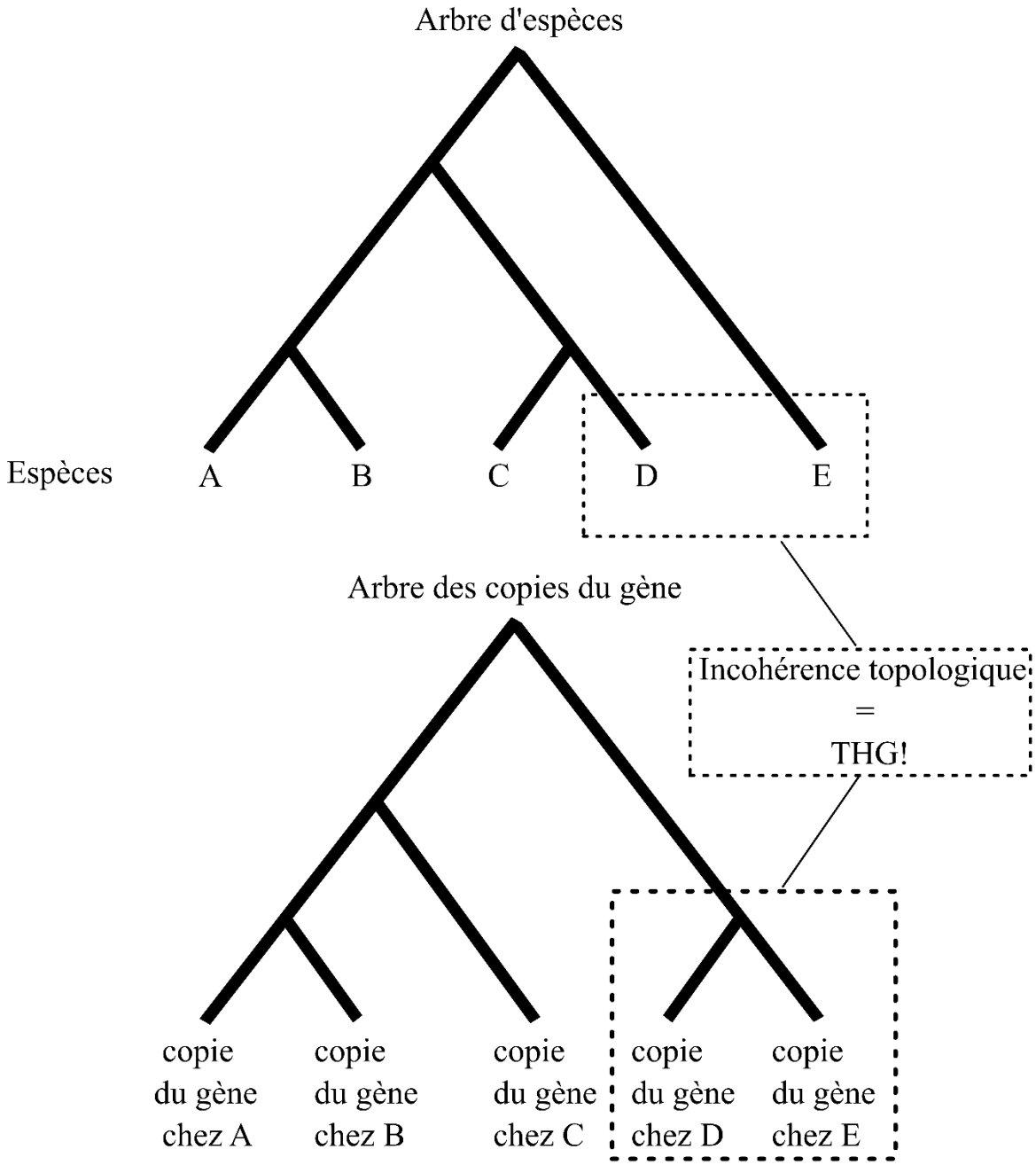


Figure 1.2 Détection de THG par une méthode phylogénétique explicite

L'arbre des espèces dont le génome contient un certain gène mobile et l'arbre du gène sont représentés respectivement de haut en bas dans cette figure. L'arbre du gène représente la distance entre les copies du gène chez les différentes espèces qui le possèdent. La présence d'une incohérence topologique entre ces 2 arbres permet d'inférer la présence d'un événement de THG.

Quant à elle, les méthodes de réconciliation cherchent à cartographier des événements évolutifs sur l'arbre de gène qui permettraient d'expliquer ses différences avec l'arbre des espèces étant donné un modèle d'évolution (Bansal, Alm, & Kellis, 2012). Plus précisément, les méthodes de réconciliation tentent d'expliquer les incohérences entre ces 2 arbres avec le minimum d'événements évolutifs possibles ou la série d'événements évolutifs associée au maximum de vraisemblance ou à l'optimisation d'un autre critère probabiliste. Les différents modèles d'évolution considèrent par exemple la duplication de gène, la perte de gène, le THG ou même la recombinaison homologue. L'avantage de la réconciliation par rapport aux autres méthodes phylogénétiques explicites est que cela fournit des informations sur les espèces impliquées dans les événements de THG et la direction du transfert. Cependant, les méthodes de réconciliation ont un temps d'exécution relativement plus élevé que les autres méthodes et sont limitées par le fait que certains événements évolutifs ont le même effet qu'une série d'autres événements évolutifs de nature différente. Par exemple, une différence topologique entre les arbres de gène et d'espèces peut autant être expliquée par le THG que par une séquence de duplication et de perte de gènes (Ravenhall et al., 2015).

1.2.2.2 Les méthodes phylogénétiques implicites

Cette catégorie de méthodes d'inférence détecte qu'un gène a été transféré horizontalement lorsque la similarité des séquences du gène est plus grande qu'attendu étant donné la distance évolutive entre les espèces qui portent ces copies du gène (**Figure 1.3 p.33**). Le principe derrière cela est expliqué par l'hypothèse de l'horloge moléculaire. Selon cette hypothèse, les gènes homologues évoluent à un rythme constant à travers différentes espèces. Cela impliquerait que, pour des gènes orthologues, la distance évolutive entre les espèces qui les possèdent serait proportionnelle à la distance entre ces séquences, ce qui n'est pas le cas pour des gènes xénologues ayant divergé par THG.

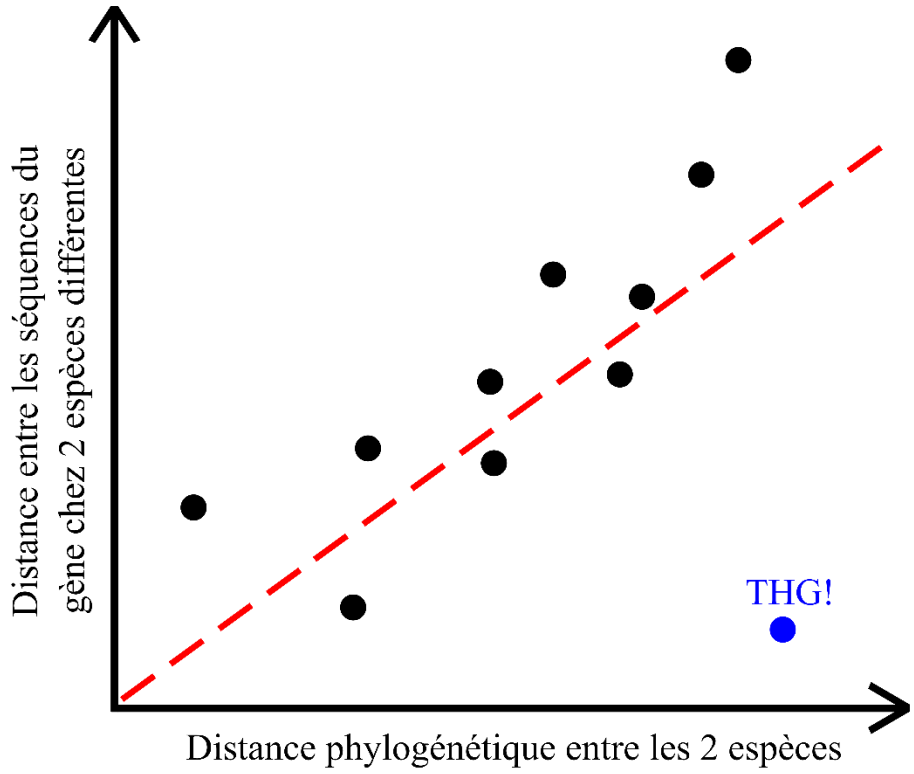


Figure 1.3 Détection de THG par une méthode phylogénétique implicite

Cette figure représente la corrélation entre la distance entre les séquences d'un gène chez 2 espèces différentes et la distance phylogénétique entre ces espèces. Chaque point représente les données pour une certaine paire d'espèces. Pour des gènes orthologues, cette corrélation devrait correspondre à une tendance linéaire positive (ligne pointillée rouge) selon l'hypothèse de l'horloge moléculaire. Certaines déviations par rapport à cette tendance attendue permettent d'inférer la présence d'un événement de THG. C'est le cas par exemple lorsque la similarité entre une paire de copies du gène est plus grande qu'attendu étant donné la similarité entre les espèces impliquées (point bleu de la figure).

Contrairement aux méthodes phylogénétiques explicites, les méthodes phylogénétiques implicites ne construisent pas d'arbres phylogénétiques, elles ont donc tendance à avoir un temps d'exécution plus court (Ravenhall et al., 2015). Cependant, elles ont aussi leurs limites. Par

exemple, la méthode des meilleures correspondances entre des espèces distantes consiste à trouver des gènes pour lesquels la similarité est élevée entre des espèces éloignées. Étant donné que l'utilisation de cette approche implique la recherche de gènes similaires entre des espèces éloignées, elle dépend fortement de la couverture taxonomique du séquençage (Ravenhall et al., 2015). De plus, tous les gènes similaires entre espèces éloignées n'ont pas nécessairement été impliqués dans un événement de THG. Certains gènes sont hautement conservés et remplissent la même fonction essentielle chez plusieurs espèces éloignées sans avoir été transférés horizontalement. Ce type de gènes serait donc des faux positifs pour cette méthode. Les gènes ribosomaux en sont un exemple (Ravenhall et al., 2015).

La méthode des meilleures correspondances entre des espèces distantes a été utilisée par Brito et ses collaborateurs pour construire une collection de gènes mobiles du microbiote intestinal humain. Avec cette méthode, cette équipe de recherche a réussi à détecter 15585 gènes mobiles provenant de 387 génomes de référence du projet de microbiome humain (HMP) et 22268 gènes mobiles provenant de 180 assemblages de génomes du projet de microbiome des communautés des îles Fidji (FijiCOMP) (Brito et al., 2016). Cette collection de gènes mobiles sera d'ailleurs celle qui sera analysée dans le cadre du présent projet. La méthode des meilleures correspondances entre des espèces distantes a pour avantage de détecter plusieurs vrais positifs et des événements de THG récents, ce qui est idéal pour étudier l'évolution pangénomique à court terme. Cependant, cette méthode peut avoir plusieurs faux négatifs à cause de sa dépendance à la couverture taxonomique du séquençage. Malgré le fait que cela est une limitation, ce n'est pas un problème d'ampleur puisque Brito et collaborateurs (2016) ont réussi à séquencer les espèces les plus abondantes du microbiote intestinal humain qui représentent la majorité de la diversité génétique microbienne dans ce milieu. À partir de cette diversité génétique, il est ensuite possible d'inférer les forces évolutives principales modulant l'évolution des gènes mobiles sur de courtes échelles de temps.

1.3 Les marqueurs de forces évolutives modulant le pangénome

Parce que la diversité nucléotidique contient des signatures de mécanismes qui modulent l'évolution des gènes et que le THG est un facteur majeur pour l'évolution du pangénome (Brito et al., 2016; McInerney et al., 2017), nous avons décidé d'analyser l'évolution des gènes mobiles avec des paramètres de la génétique des populations. Les paramètres que nous avons choisis pour détecter les forces majeures agissant sur les gènes mobiles sont :

(1) θ_π, l'estimateur de diversité nucléotidique calculé à partir du nombre moyen de différences par paire de séquences :

$$\theta_\pi = \frac{k}{\langle {n \atop 2} \rangle}$$

où k est le nombre de différences total entre les paires de séquences, n représente le nombre séquences du gène alignées et $\langle {n \atop 2} \rangle$ représente le nombre de paires de séquences comparées. Cet estimateur est sensible aux mutations à fréquence intermédiaire puisque celles-ci maximisent le nombre de différences entre les paires de séquences (Tajima, 1989).

(2) θ_w, l'estimateur de diversité nucléotidique calculé à partir du nombre de sites polymorphiques :

$$\theta_w = \frac{S}{a1}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

où S est le nombre de sites polymorphiques et a1 est un facteur de normalisation qui reflète la taille de l'échantillonnage. Cet estimateur est sensible aux mutations à faible fréquence présentes à travers différents sites polymorphiques (Tajima, 1989).

(3) le *D de Tajima*, qui mesure la différence entre θ_π et θ_w, et donc l'équilibre entre les mutations à fréquence intermédiaire et les mutations à faible fréquence :

$$D_{Tajima} = \frac{\theta_\pi - \theta_w}{\sqrt{\widehat{Var}(\theta_\pi - \theta_w)}}$$

où $\widehat{Var}$ représente la variance attendue de $(\theta_\pi - \theta_w)$ (Tajima, 1989).

(4) *dN/dS*, le ratio entre le taux de substitutions non synonymes et le taux de substitutions synonymes :

$$\frac{\widehat{dN}}{dS} = \frac{Nb_{nsm}/Nb_{nss}}{Nb_{sm}/Nb_{ss}}$$

où Nb_{nsm} est le nombre de substitutions non synonymes, Nb_{nss} est le nombre de sites non synonymes, Nb_{sm}est le nombre de substitutions synonymes, et Nb_{ss} est le nombre de sites synonymes.

On note que θ_π et θ_w sont différents estimateurs de la diversité nucléotidique (θ) d'un gène et que celle-ci peut aussi être estimée par la taille efficace de la population (N_e), qui mesure la diversité génétique et l'efficacité de la sélection dans une population :

$$\theta = N_e * \mu$$

Où μ est le taux de mutation

On note aussi que des valeurs positives de D de Tajima révèle un excès de mutations à fréquence intermédiaire par rapport à un modèle neutre d'évolution dans lequel la taille de la population dans lequel le gène évolue est constante. Cela peut être expliqué par plusieurs forces évolutives comme la réduction de la taille de la population ou la sélection purificatrice. À l'opposé, des valeurs négatives de D de Tajima révèlent un excès de mutations à faible fréquence par rapport à un modèle neutre d'évolution où la taille de la population dans laquelle le gène évolue est constante. Cela peut être expliqué par l'augmentation de la taille de la population ou la sélection positive récente. Pour ce qui est de dN/dS, des valeurs négatives reflètent la prédominance de la sélection négative tandis que des valeurs positives reflètent la prédominance de la sélection positive.

Ces paramètres sont habituellement estimés à partir de l'alignement et de l'arbre des allèles d'un gène dans une population. Nous expliquerons à la section 3.5 (**p.87**) comment nous avons estimé ces paramètres à l'aide de notre jeu de données de lectures de séquençage métagénomique. Ces paramètres de la génétique des populations permettront de caractériser l'équilibre entre la sélection, la dérive génétique et le transfert horizontal, ce qui est le sujet principal des modèles d'évolution pangénomique.

1.4 Modèles d'évolution pangénomique

Ces modèles permettent de déterminer les forces évolutives prédominantes modulant la variation du contenu en gène des procaryotes et l'évolution des séquences de gènes mobiles. Ils se divisent en 3 catégories, soit les modèles adaptatifs, les modèles neutres et les modèles presque neutres.

1.4.1 Les modèles adaptatifs

Les modèles adaptatifs d'évolution pangénomique considèrent que les pangénomes et le transfert horizontal sont adaptatifs (McInerney et al., 2017; Sela et al., 2016). Cela implique que la sélection a plus d'effets que la dérive génétique sur l'évolution des gènes mobiles et du contenu en gène des procaryotes. Des publications récentes illustrent bien les arguments en faveur du modèle adaptatif d'évolution pangénomique.

Tout d'abord, Sela et collaborateurs (2016) ont développé un modèle adaptatif expliquant l'évolution de la taille des génomes des procaryotes en utilisant les données génomiques de 707 espèces procaryotes relativement distantes. Ils ont tout d'abord remarqué que la taille des génomes de ces espèces corrèle négativement avec dN/dS, qui mesure la force et le type de sélection en action. Cela est cohérent avec le fait que le gain de gène a des effets adaptatifs sur les procaryotes. Cela s'explique par le fait que la taille des génomes de ces espèces corrèle positivement avec N_e, qui mesure l'efficacité de la sélection. Enfin, cette équipe de recherche a développé un modèle d'évolution qui inclut le gain de gène, la perte de gène et leurs effets sélectifs. Dans leur modèle, le gain et la perte de gènes maintiennent l'équilibre de la taille du génome et ont des effets sélectifs opposés. Le modèle tient également compte de la taille efficace de la population. À partir de simulations de ce modèle, ils ont constaté qu'un scénario dans lequel le gain de gène est, en moyenne, légèrement bénéfique, explique le mieux les données de taille du génome et de diversité nucléotidique des 707 génomes procaryotes à l'étude.

En se basant sur une synthèse de données génomiques microbiennes et de modèles comprenant celui de Sela et de ses collaborateurs (2016), un autre groupe, composé de McInerney et ses collaborateurs (2017), a aussi proposé qu'un modèle adaptatif explique le mieux l'évolution des pangénomes. En effet, ceux-ci ont observé que les pangénomes plus larges ont tendance à apparaître dans des espèces avec un N_e plus grand à cause des effets bénéfiques du gain de gène, d'une plus grande efficacité de la sélection et d'un grand nombre de niches écologiques disponibles pour ces espèces.

1.4.2 Les modèles neutres

Contrairement aux modèles adaptatifs, les modèles neutres considèrent que la dérive génétique a plus d'effets que la sélection sur l'évolution des gènes mobiles et du contenu en gène des procaryotes. La dichotomie entre ces modèles neutres et les modèles adaptatifs est encore

d'actualité dans la littérature scientifique. Plusieurs publications contiennent des observations génomiques qui concordent avec un modèle non adaptatif d'évolution pangénomique. Par exemple, en analysant les données génomiques de 90 espèces procaryotes, Andreani et ses collaborateurs (2017) ont observé que la fluidité du génome de ces espèces, définie par le ratio entre le nombre de familles de gènes uniques contenues et le nombre moyen de familles de gènes entre des paires de génomes aléatoires, corrèle positivement avec la diversité nucléotidique synonyme. Cette observation n'exclut pas le rôle de la sélection, mais s'explique le plus parcimonieusement par un modèle neutre d'évolution pangénomique. En effet, puisque la diversité nucléotidique synonyme corrèle aussi positivement avec N_e, on s'attend à ce que des espèces ayant un plus grand N_e aient des génomes plus divers et fluides.

1.4.3 Les modèles presque neutres

Malgré l'opposition entre les modèles adaptatifs et neutres, il est possible de réconcilier ces modèles avec des interprétations plus nuancées de l'évolution pangénomique. C'est le cas des modèles d'évolution presque neutres (quasi neutres) selon lesquelles l'équilibre entre la sélection et la dérive génétique dépend de la taille efficace de population. Plus précisément, parce que le coefficient de sélection des gènes mobiles est en moyenne faible ou presque neutre, ceux-ci peuvent ne pas être capable d'échapper les effets de la dérive génétique comme la perte de gène ou la fixation aléatoire de mutations délétères lorsque N_e, qui mesure l'efficacité de la sélection, est petit. Par exemple, en se basant sur les données de 153 espèces procaryotes, Bobay et Ochman (2018) ont tout d'abord observé que, la plupart du temps, le taux de remplacement des gènes dans ces génomes ne corrèle pas significativement avec *dN/dS*, ce qui est cohérent avec le fait que les gènes mobiles seraient en moyenne presque neutres. En revanche, tout comme McInerney et ses collaborateurs (2017), ils ont observé que la taille des pangénomes de ces espèces corrèle positivement avec N_e et attribue cela à une augmentation de l'efficacité de la sélection. Ces deux observations peuvent sembler contradictoires puisqu'elles appuient respectivement un modèle neutre et un modèle adaptatif. Cependant, Bobay et Ochman expliquent cela par un modèle presque neutre d'évolution soit le modèle barrière-dérive. Selon ce modèle, puisque le coefficient de sélection des gènes mobiles est en moyenne presque neutre, ceux-ci peuvent sembler virtuellement neutres lorsque N_e est petit. Par contre, lorsque la taille efficace de population augmente, ces gènes peuvent échapper aux effets de la dérive génétique et être maintenus dans les pangénomes de ces espèces, ce qui agrandit la taille de ceux-ci. Le nom du modèle barrière-dérive fait référence au

seuil de N_e à partir duquel les gènes mobiles peuvent échapper aux effets de la dérive génétique comme la perte de gène ou à l'accumulation de mutations délétères.

1.5 Outils bio-informatiques

Afin d'inférer le type de modèle qui décrit le mieux l'évolution pangénomique durant de courtes échelles de temps, l'analyse des données de séquençage métagénomique et la validation du modèle par des simulations seront nécessaires.

1.5.1 Analyse des données métagénomiques avec anvi'o

Anvi'o est une plateforme informatique constitué d'outils permettant la visualisation et l'analyse avancée de données omiques (Eren et al., 2015). Cette plateforme permet d'assigner des lectures de séquençage métagénomique à leur espèce d'origine, de caractériser la variation nucléotidique des gènes, d'étudier les pangénomes bactériens, de prédire le nombre de génomes bactériens dans un échantillon à partir des lectures de séquençage métagénomique ou même de détecter la présence d'une contamination dans un échantillon microbien. Dans le cadre de ce projet, nous utiliserons l'outil d'identification de variantes de séquence d'anvi'o afin de détecter les mutations des gènes mobiles du microbiote intestinal humain à partir de l'alignement entre les séquences de références de ces gènes et les lectures de séquençage métagénomique. L'outil d'identification des variantes de séquence d'anvi'o a comme avantage de simplifier la représentation et l'analyse des données métagénomiques, de permettre de filtrer ces données afin de diminuer l'impact des erreurs de séquençage et de pouvoir s'exécuter rapidement puisqu'il permet de faire du calcul de haute performance, notamment grâce à l'utilisation de plusieurs cœurs de calcul. À partir des résultats de cette analyse, il sera ensuite possible de caractériser la diversité nucléotidique des gènes mobiles et l'évolution du pangénome grâce aux paramètres de génétique des populations mentionnés à la section 1.3.

1.5.2 Simulation évolutive avec SodaPop

Afin de valider les conclusions tirées de l'analyse des données métagénomiques réelles et afin de mieux interpréter nos observations, nous exécuterons des simulations évolutives microbiennes grâce à SodaPop (Gauthier, Di Franco, & Serohijos, 2019). SodaPop est une suite logicielle libre d'accès qui permet de simuler l'évolution de populations de cellules asexuées basée sur des paysages adaptatifs de protéines (Gauthier, Di Franco, & Serohijos, 2017). Les paysages

adaptatifs de protéines sont des modèles mathématiques pour l'estimation du succès reproducteur des cellules, basés sur les caractéristiques des protéines. Par exemple, les paysages adaptatifs protéiques courants sont les paysages de repliement des protéines ou les paysages d'activités enzymatiques. Les paysages de repliement des protéines supposent qu'il existe une corrélation positive entre le succès reproducteur des cellules et la stabilité du repliement des protéines, estimée par la variation d'énergie libre de repliement ($\Delta\Delta G$) par exemple (Gauthier et al., 2019). En d'autres termes, plus la cellule a des protéines actives stables, plus son succès reproductif a tendance à augmenter. Quant aux paysages d'activité enzymatique, ils supposent qu'il existe une corrélation positive entre le succès reproducteur des cellules et l'activité des enzymes, estimée avec l'efficacité catalytique, c'est-à-dire le rapport entre la constante catalytique (K_{cat}) et la constante de Michaelis (K_M) (Berg, Tymoczko, & Stryer, 2002; Gauthier et al., 2019).

Durant les simulations, SodaPop considère différents aspects biologiques tels que la taille de la population, les séquences génétiques, les mutations, les forces évolutives et les contraintes biophysiques des protéines. En considérant ces facteurs en plus d'un paysage adaptatif de protéines, SodaPop cherche à évaluer l'influence des mutations sur les caractéristiques biophysiques des protéines, à suivre l'évolution des lignées microbiennes, à modéliser la fixation des mutations et à déterminer les effets des contraintes biophysiques des protéines et des paramètres génétiques des populations sur l'évolution (Gauthier et al., 2019). C’est donc une approche intégrative originale pour étudier l’évolution des procaryotes. Les étapes d’exécution de SodaPop sont illustrées à la **figure 1.4 (p.41)**.

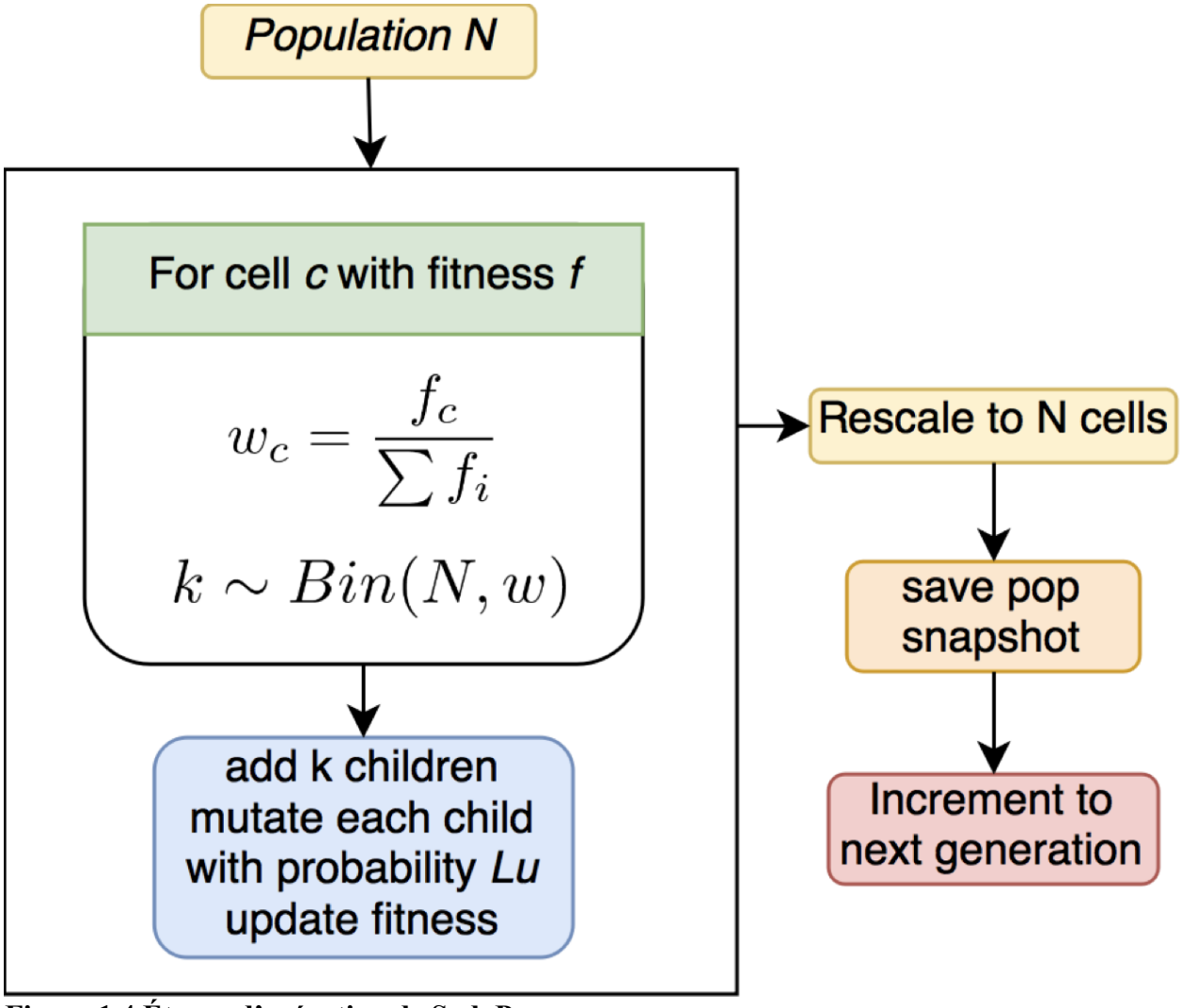


Figure 1.4 Étapes d'exécution de SodaPop

Lors de chaque étape de simulation, SodaPop démarre avec une population de N cellules dont le contenu en gène dépend de l'espèce de ces cellules. Pour chaque cellule de la communauté microbienne simulée, SodaPop calcule le succès reproducteur en fonction des paysages adaptatifs des protéines et des séquences de gènes. Les valeurs sélectives (ω_c) sont déterminées à partir de la valeur de succès reproducteur relative et le nombre de descendants (k) de chaque cellule est déterminé avec une distribution binomiale en fonction de la taille de la population et des valeurs sélectives (Gauthier et al., 2019):

$$\omega_c = \frac{f_c}{\sum_1^N f_i}$$

$$k = B(N, \omega_c)$$

où ω_c est le coefficient de sélection de la cellule c, f_i est la valeur de succès reproducteur de la cellule i, B réfère à la distribution binomiale, N est la taille de la population et k est le nombre de descendants de la cellule c. Ensuite, chaque cellule de la nouvelle génération est mutée avec une probabilité Lμ, où L est la longueur du génome et μ est le taux de mutation. Les dernières étapes de l'itération consistent à ajuster la taille de la population à N, à sauvegarder les données et à passer à l'itération suivante (Gauthier et al., 2019).

Dans le cadre de ce projet, nous ajouterons un module de simulation du THG et de la perte de gènes afin d'étudier l'évolution pangénomique à court terme.

2 Objectifs et hypothèses

Tout d'abord, ce projet cherche à déterminer quel modèle d'évolution explique le mieux l'évolution pangénomique à court terme. Puisque le THG semble être en moyenne légèrement bénéfique (Bobay & Ochman, 2018; Sela et al., 2016) et que certaines familles de gènes mobiles comme les gènes de résistance aux antibiotiques sont sélectionnés dans le microbiote intestinal humain durant le temps de vie d'un individu (Jiang et al., 2019), nous devrions être capables d'observer qu'un modèle adaptatif d'évolution explique le mieux l'évolution pangénomique sur de courtes échelles de temps. Ainsi, ma première hypothèse est que :

1. La mobilité des gènes devrait être associée à des signatures de sélection positive. Plus précisément, il devrait y avoir une corrélation significative entre des marqueurs de sélection, comme le D de Tajima ou *dN/dS*, et le taux de THG d'un gène ou son niveau de mobilité même sur de courtes échelles de temps.

Une hypothèse alternative serait que les pressions de sélection sont effectives que pour certains gènes sur de courtes échelles de temps.

Enfin, ce projet cherche à évaluer l'impact des attributs de l'hôte du microbiote intestinal sur l'évolution pangénomique durant de courtes échelles de temps évolutif. En comparant des populations humaines ayant divergé il y a des milliers d'années, Brito et ses collaborateurs (2016) ont découvert que la composition en espèces et le réservoir de fonctions de gènes mobiles du microbiote intestinal étaient significativement influencés par des attributs comme la diète de l'hôte et son réseau social. En revanche, il n'est pas garanti que l'effet de ces facteurs soit perceptible sur de courtes échelles de temps, notamment en comparant des individus d'une même population qui ont souvent où un mode de vie et un réseau social similaires. Je m'attends donc à observer une corrélation faible entre les attributs d'hôte et la diversité nucléotidique des gènes mobiles sur de courtes échelles de temps. Au lieu d'être liées à l'hôte humain, les pressions de sélection seront plutôt liées à la fonction des gènes chez les espèces microbiennes. Donc, ma deuxième hypothèse est que :

2. La fonction du gène a plus d'impact que les attributs de l'hôte sur l'évolution des gènes mobiles.

3 Article

Pangenome sequence evolution within human gut microbiomes is explained by gene-specific rather than host-specific selective pressures

Arnaud N'Guessan[1,2], Ilana Lauren Brito[3], Adrian W.R. Serohijos[1,2 *], B. Jesse Shapiro[4,5,6 *]

[1]*Departement de Biochimie,* [2]*Centre Robert-Cedergren en Bio-informatique et Génomique, Université de Montréal, 2900 Édouard-Montpetit, Montréal, Québec H3T 1J4, Canada*

[3]*Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA*

[4]*Département de sciences biologiques, Complexe des sciences, Université de Montréal, 1375 Avenue Thérèse-Lavoie-Roux, Montréal, QC H2V 0B35*

[5]*Department of Microbiology and Immunology, McGill University, Montreal, QC, Canada*

[6]*McGill Genome Centre, Montreal, QC, Canada*

Target journal: Nature Microbiology

[*]*Correspondence: adrian.serohijos@umontreal.ca, jesse.shapiro@mcgill.ca*

Author contributions:

For this publication, Arnaud N'Guessan implemented all the data analysis methods and the simulation model. He also wrote the initial manuscript and created the figures. Dr. Ilana Brito commented the manuscript and facilitated the access to the metagenomic data and sample metadata. Finally, Dr. Serohijos and Dr. Shapiro assisted in the interpretation of the results and commented on the data analysis methods, manuscript and figures to allow their improvement.

Keywords: Pangenome, Evolution, Mobile genes, Horizontal Gene Transfer, Human gut microbiome, Evolutionary Simulations

3.1 Abstract

Pangenomes – the cumulative set of genes encoded by a species – arise from evolutionary forces including horizontal gene transfer (HGT), drift, and selection. The relative importance of drift and selection in shaping pangenome structure has been recently debated, and the role of sequence evolution (point mutations) within mobile genes has been largely ignored, with studies focusing mainly on patterns of gene presence or absence. The effects of drift, selection, and HGT on pangenome evolution likely depends on the time scale being studied, ranging from ancient (*e.g.,* between distantly related species) to recent (*e.g.*, within a single animal host), and the unit of selection being considered (*e.g.,* the gene, whole genome, microbial species, or human host). To shed light on pangenome evolution within microbiomes on relatively recent time scales, we investigate the selective pressures acting on mobile genes using a dataset that previously identified such genes in the gut metagenomes of 176 Fiji islanders. We mapped the metagenomic reads to mobile genes to call single nucleotide variants (SNVs) and calculate population genetic metrics that allowed us to infer deviations from a neutral evolutionary model. We found that mobile gene sequence evolution varied more by gene family than by human social attributes, such as household or village membership, suggesting that selection at the level of gene function is most relevant on these short time scales. Patterns of mobile gene sequence evolution could be qualitatively recapitulated with a simple evolutionary simulation, without the need to invoke an adaptive advantage of mobile genes to their bacterial host genome. This suggests that, at least on short time scales, a majority of the pangenome need not be adaptive. On the other hand, a subset of gene functions including defense mechanisms and secondary metabolism showed an aberrant pattern of molecular evolution, consistent with species-specific selective pressures or negative frequency-dependent selection not seen in prophages, transposons, or other gene categories. That mobile genes of different functions behave so differently suggests stronger selection at the gene level, rather than at the genome level. While pangenomes may be largely adaptive to their bacterial hosts on longer evolution time scales, here we show that, selection acts on individual genes, but not in a way that is necessarily adaptive to human host or microbial host cell fitness.

3.2 Introduction

Human gut microbial communities (or microbiomes) impact diverse aspects of human health, such as food digestion, nutritional uptake, immunity, and inflammation (Brito et al., 2016; Valdes et al., 2018). The gut microbiome is shaped by both ecological factors, such as shifts in species abundance or strain replacements, and evolutionary forces, such as mutation, horizontal gene transfer (HGT), drift and selection (Garud & Pollard, 2020). In particular, microbes in the gut dynamically and frequently exchange genetic material through HGT (Vos, Hesselman, Te Beek, van Passel, & Eyre-Walker, 2015), resulting in pangenomes (the total set of genes observed in all members of a species or population) which are often much larger than an individual genome size (Jiang et al., 2019; McInerney et al., 2017; Sela et al., 2016). Some studies have shown that horizontally transferred (mobile) genes could contribute to environmental adaptation, notably through the propagation of antibiotic resistance (Jiang et al., 2019). However, there are contexts in which pangenome evolution could be driven more by drift than by selection. For instance, the evolution of endosymbionts or intracellular pathogens, which have small effective population sizes, is generally driven by drift, resulting in small pangenomes (Giovannoni, Cameron Thrash, & Temperton, 2014). In contrast, selection seems to play a bigger role in free-living microbes, like hydrothermal vent bacteria (Moulana, Anderson, Fortunato, & Huber, 2020). Whether pangenome evolution is mainly driven by selection (an adaptive model) or drift (a non-adaptive or neutral model) is a question that has generated some controversy (Andreani et al., 2017; Bobay & Ochman, 2018; McInerney et al., 2017; Sela et al., 2016).

Answering this question depends on the time scale being studied. For example, long-term evolution (*e.g.* among distantly related species or among all extant members of a species) versus near-term evolution (*e.g.* among a locally coexisting population of a species) may experience different regimes of drift and selection. On long time scales, using data from distantly related genomes that diverged millions of years ago (McInerney et al., 2017; Sela et al., 2016), and at the whole-genome scale, adaptive and non-adaptive models have been proposed and are still a source of contention. A model in which gene gain by HGT is predominantly adaptive provides a good fit to distantly related genomes from the NCBI database (Sela et al., 2016). In that work, Sela and collaborators developed a model of prokaryotic genome size evolution that includes gene gain, gene loss, and their fitness effects (Brito et al., 2016). In their model, gene gain and loss maintain genome size equilibrium and have opposite fitness effects. The model also accounts for species

effective population size (N_e), which measures genetic diversity and effectiveness of selection in a population, and is dependent on census population size and its fluctuations (Bobay & Ochman, 2018) as well as on varying intensities of purifying, positive, or fluctuating natural selection. From simulations of this model, they found that a scenario in which gene gain is, on average, slightly beneficial best fits genome size and nucleotide diversity data from 707 prokaryotic genomes. Based on a synthesis of population genomic data and models including Sela and collaborators' model (Sela et al., 2016), another group led by McInerney and collaborators argued that an adaptive model best explains pangenome evolution because more diverse pangenomes tend to arise in species with larger N_e due to beneficial gene gain, higher efficacy of selection, and a large number of micro-niches available to the species (McInerney et al., 2017).

In contrast, Andreani and collaborators (2017) observed that genome fluidity, defined as the ratio between the number of unique gene families and the average number of gene families between random genome pairs, significantly correlates with synonymous nucleotide diversity in 90 bacterial species. Although this does not exclude a role for selection, the observation is most parsimoniously explained by a neutral model. Similarly, Bobay and Ochman (2018) observed that gene turnover does not significantly correlate with *dN/dS*, which measures selection on protein-coding genes. They also found that N_e correlates positively with pangenome size for most of the 153 analyzed prokaryotic species. Similar to McInerney and collaborators, they attributed this to an increased effectiveness of selection in species with larger N_e and that most of the accessory genes, those that are present in some but not all strains of a species, are slightly beneficial (McInerney et al., 2017). The fact that Bobay and Ochman (2018) found evidence for both adaptive and neutral pangenome evolution may seem contradictory. However, they reconciled these observations by proposing a nearly neutral model of drift-barrier evolution. This model describes the balance between selection and drift. More precisely, it assumes that most accessory genes in the pangenome are slightly beneficial, such that they can be considered neutral when N_e is small, but they can escape the effects of drift and spread when the selective coefficient s exceeds $1/N_e$.

Resolving the balance of evolutionary forces influencing pangenomes also depends on the biological scale or unit of evolution. For example, the consequences of selection at the level of single genes, whole genomes, microbial species or human hosts could yield different patterns. The studies above focused on adaptation at the whole-genome level, but selection also acts at the level

of individual genes (Moulana et al., 2020; Shapiro, 2017; Takeuchi, Cordero, Koonin, & Kaneko, 2015). Mobile genes in particular may have their own N_e, which could be distinct from the N_e of the whole genome of a species (Shapiro, 2017). For example, there is an entire class of mobile genes, including phage and other "selfish" elements that have effectively instantaneous HGT rates (Wolf, Makarova, Lobkovsky, & Koonin, 2016). Other mobile genes may provide rapid adaptive value to their bacterial hosts, such as in the gut microbiome of humans with different diets or lifestyles (Brito et al., 2016). Therefore, based on their patterns of presence or absence, some mobile genes appear to be selected to favour their own replication (selfish) while others may provide benefits to their bacterial or even human hosts (Hehemann et al., 2010).

All the studies above investigated pangenome evolution among distantly related genomes over relatively ancient time scales. Yet selective pressures might differ on recent and shorter evolutionary time scales, such as within local populations of bacteria over dozens rather than millions of years. However, a targeted investigation of the population genetics of mobile genes on short time scales is still missing. To study pangenome evolution on shorter evolutionary time scales and at the level of individual genes, we used a dataset from Brito and collaborators composed of 37 853 mobile genes involved in recent HGT events in the human gut (Brito et al., 2016). We mapped metagenomic reads from a cohort of 176 Fiji islander gut microbiomes to this set of mobile genes. From the mapped reads, we identified single nucleotide variants (SNVs) segregating within microbiomes, from which we calculated population genetic metrics such as *dN/dS* and *Tajima's D* that contain information about evolutionary and demographic history of mobile genes. In contrast to studies over longer evolutionary time scales, which have concluded that pangenome evolution is largely adaptive, we find that many aspects of pangenome molecular evolution on shorter time scales can be explained without invoking any adaptive benefit of mobile genes to their human hosts. However, a small subset of genes with distinct functions show dramatically different signature of molecular evolution, suggesting that selection acts at the level of gene function. Our results suggest that while host-related selective pressures may be strong over long evolutionary time scales, selection at the level of individual genes might predominate over shorter human time scales.

3.3 Results and discussion

3.3.1 Gene mobility correlates positively but not strongly with metagenomic coverage

To study pangenome evolution on time scales on the order of a human lifespan, we used an existing collection of mobile genes identified in 387 isolate genomes from the Human Microbiome Project (HMP) and 180 single-cell genomes from the Fiji Community Microbiome Project (FijiCOMP). Selected single-cell genomes came from 31 different genera and had less than 10% putative contamination called by CheckM (Brito et al., 2016; Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015). The mobile genes were identified in genomic regions containing at least 500bp with >99% nucleotide identity over >50% of their sequence length between distantly related single-cell bacterial genomes (<97% identity in 16S rRNA), suggesting that HGT occurred within an individual human gut microbiome (Brito et al., 2016). Ribosomal genes, which tend to be highly conserved, were excluded from this set of mobile genes as they could represent false-positive HGT events (Brito et al., 2016). This procedure is strict, yielding likely true positive HGT events, at the expense of many false negatives (Brito et al., 2016; Smillie et al., 2011). We considered only genes with at least 10X metagenomic sequence coverage, and only metagenomes with at least 500 genes passing this coverage threshold. These filters yielded a total of 7,990 mobile genes out of the 37 853 genes present in the original dataset, and 175 out of 176 metagenomes, each from a different person from Fiji. We operationally defined gene mobility as the number of single-cell genomes in which a mobile gene was found. Gene mobility ranged from 1-16 species (mean = 2.73, standard deviation = 2.42; **Figure 3.S1 p.51**) and is probably an underestimate of the true HGT rate because it was estimated from a limited sample (180 genomes) of the diversity in Fijian islanders' gut. This could also be explained by small or incomplete assemblies of the single-cell genomes. Nonetheless, this dataset provides allows us to assess the balance of evolutionary forces in the pangenome on short timescales.

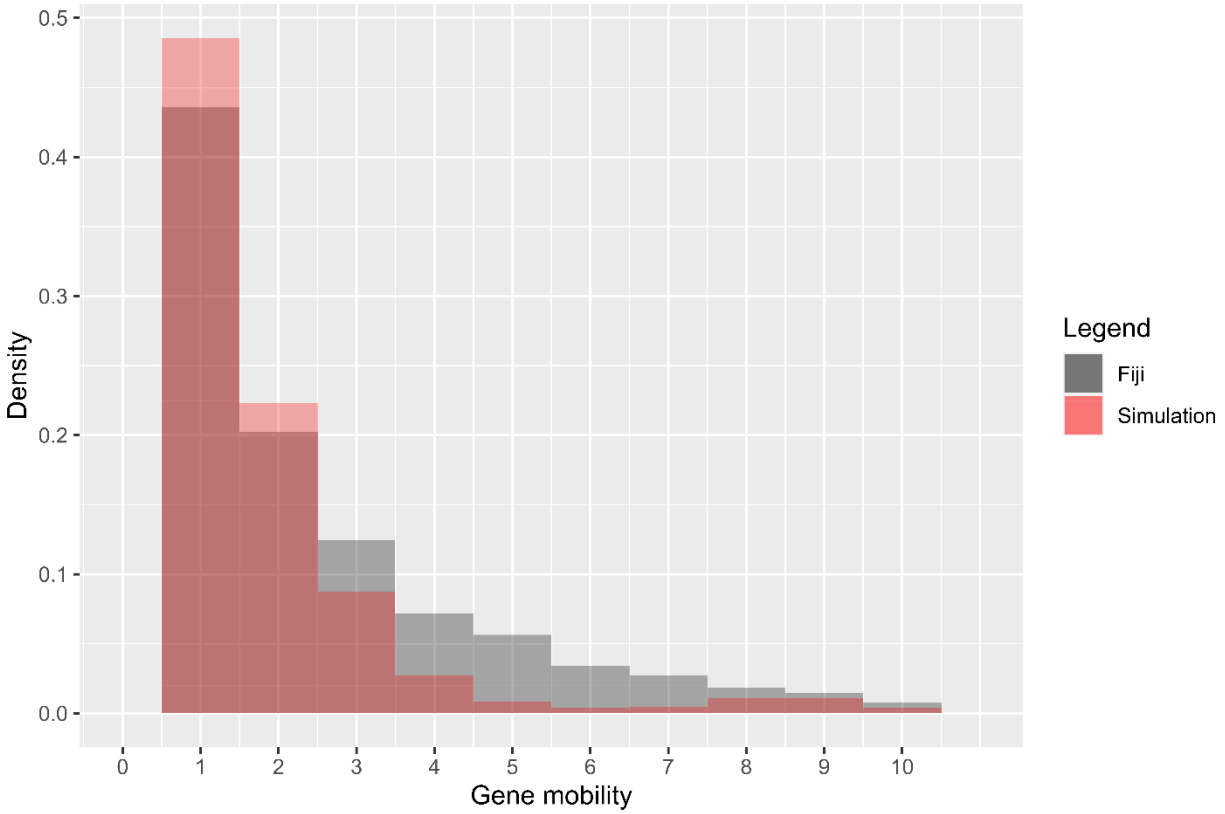


Figure 3.S1. Gene mobility distribution in simulation vs in Fiji dataset

Simulations in which HGT is slightly adaptive and the HGT selection coefficient is exponentially distributed produce a gene mobility distribution with a similar shape than the one observed in the Fiji dataset. However, their quantitative similarity is not significant (Kolmogorov-Smirnov test p-value < 0.05) and the range of mobility in simulation is [1,10] while it is [1,16] in the Fiji dataset so the distribution is truncated here to a maximum of 10 species. The simulation presented (red) included 5000 cells, 10 species, 500 genes per cells at equilibrium, a simulation time of 10^5 generations, HGT rate = 10μ and HGT selection coefficient parameter λ = 1E5 (HGT is slightly adaptive; s = 1E-5 in average).

We began by asking whether our mobility metric behaves as expected in quantifying the spread of mobile genes in the gut. Assuming that genes with higher mobility will occur in more species, we expect them to be more deeply covered by metagenomic sequence reads. Consistent with this expectation, we found that a gene's mobility is positively correlated with its depth of metagenomic read coverage (**Figure 3.1 p.52 and Table 3.S1 p.54**). The expectation of a positive

correlation is not guaranteed because some mobile genes, such as selfish elements, have deleterious effects(Vogan & Higgs, 2011) and can be subject to negative frequency-dependent selection (Corander et al., 2017; Domingo-Sananes & McInerney, 2019; Takeuchi et al., 2015) such that they are carried only by a fraction of individuals within a species, even if prevalent across species. The correlation between gene mobility and coverage is significantly positive in 169 out of 175 gut metagenomes (Bonferroni-adjusted p-value < 2.2×10^{-16}), but the adjusted R^2 and slope values are relatively modest (**Figure 3.1 p.52 and Figure 3.S2 p.53**). Varying selective pressures across mobile genes (*e.g.* deleterious effects and negative frequency-dependent selection) might be responsible for reducing the scaling between gene mobility and coverage, but not enough to flatten the relationship completely. We conclude that gene mobility, even if estimated from a relatively small sample of 180 gut bacterial genomes, behaves approximately as expected: generally leading to higher gene copy numbers.

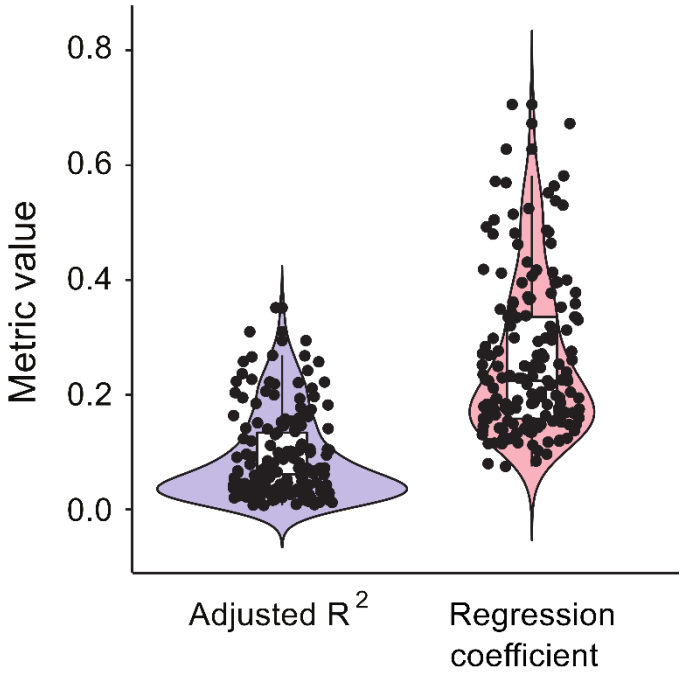


Figure 3.1 The correlation between gene mobility and metagenomic sequencing coverage is positive but widely variable

The boxplots and violin plots show the distributions of adjusted R^2 values (blue) and slopes (red) across samples (individuals from Fiji) for the correlation between coverage (average depth per site) and gene mobility. The black dots represent the 169 samples (out of 175 tested) in which the

correlation is significant (*t*-test, Bonferroni-adjusted p-value < 0.05; Methods). Examples of this correlation in four randomly selected samples are shown in **Figure 3.S2 (p.53)**.

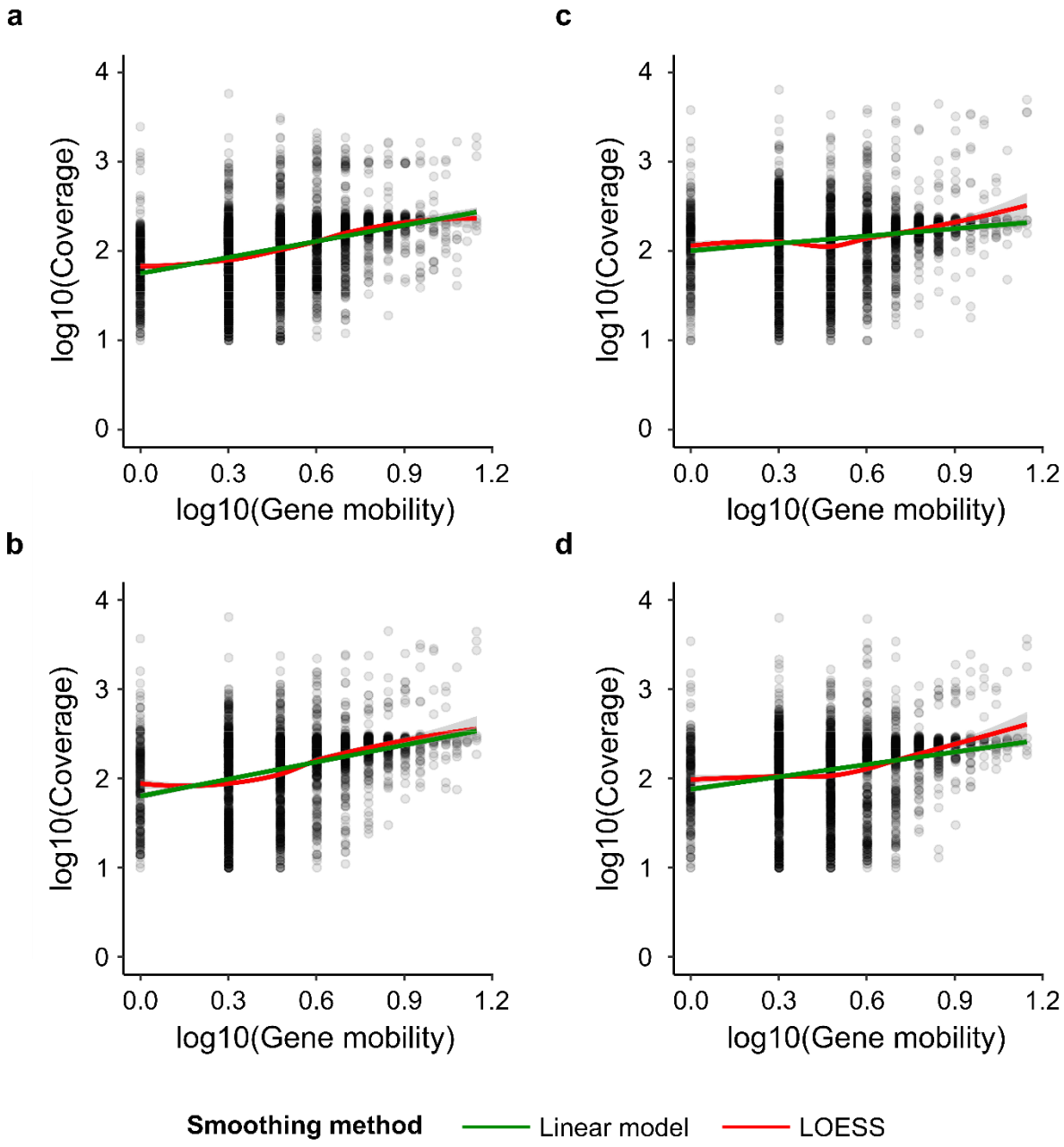


Figure 3.S2 Coverage in function of gene mobility in log10 scale across 4 samples

These figures show examples of the correlation between a mobile gene coverage and its mobility in log10 scale for samples a) G30512, b) G30771, c) G30520 and d) G30804. The correlation

observed in these samples is positive, which is consistent with the main trends. Additionally, coverage and mobility range are shown in the figure. Coverage is always greater than 10x, which is a minimal requirement for variant calling. Finally, linear model (green) and LOESS (red) smoothing curves are represented by the colored lines.

Model	Description	R^2	Log-likelihood	LRT p-value (Nested model vs M1 model)
M1	Coverage vs Mobility + all random factors, i.e. Sample and COG category	0.35	-317 537	-
M2	Coverage vs Mobility + COG category (without Sample)	0.12	-355 586	$< 2.2x10^{-16}$
M3	Coverage vs Mobility + Sample (without COG category)	0.31	-321 438	$< 2.2x10^{-16}$

Table 3.S1 Regression strength of the linear mixed model $\boldsymbol{Coverage \sim Mobility + Sample + COG\ category}$ and nested models LRT

Because the likelihood ratio test (LRT) p-value is highly significant for each of the nested models, all the random effects tested, i.e. sample and COG category, have a significant effect on the regression. The M3 model R^2 is much higher than the M2 model R^2, suggesting that the relationship between Coverage and Mobility varies more across samples than across COG categories.

3.3.2 Estimating population genetic metrics from metagenomic data

The relationship between metagenomic coverage and gene mobility is generally positive but varies substantially across individuals (**Figure 3.1 p.52**). We therefore sought to ask whether this variation could be explained by either gene-specific factors (*e.g.* gene mobility and COG functional categories) or human-specific factors, such age, diet, or social networks (Brito et al., 2016; Tatusov, Galperin, Natale, & Koonin, 2000). Both gene-specific and human-specific factors are known to influence the patterns of mobile gene presence/absence across bacterial genomes (Wolf et al., 2016) and human hosts (Brito et al., 2016; Cordero & Polz, 2014; Gardon, Biderre-Petit, Jouan-Dufournel, & Bronner, 2020; Koonin & Wolf, 2010), yet it is unclear if these patterns are explained

by selection or drift. Here, we used the tools of population genetics to study molecular evolution of mobile genes based on their patterns of single nucleotide variants (SNVs) segregating in gut metagenomes. We quantified mobile gene sequence evolution using four population genetic metrics that detect selection and capture deviations from a neutral evolutionary model:

(1) θ_π, the nucleotide diversity calculated from the average number of pairwise nucleotide differences among metagenomic reads,

(2) θ_w, the nucleotide diversity calculated from the normalized number of segregating/polymorphic sites in metagenomic reads,

(3) *Tajima's D*, the normalized difference between θ_π and θ_w, and

(4) *dN/dS*, the ratio of nonsynonymous to synonymous substitution rates, measuring selective constraints at the protein level.

We note that our estimate of *dN/dS,* based on mapping metagenomic reads that could come from the same or different species, is a mixture of within-species polymorphism (often called *pN/pS*) and between-species divergence (*dN/dS*), but we refer to this hybrid metric as *dN/dS* for simplicity. We further note that θ_π and θ_w are two different estimators of the population mutation rate, $\theta = 2N_e\mu$, where μ is the mutation rate and N_e is the effective population size. This difference in the two estimators is captured by *Tajima's D*. In particular, *Tajima's D<0* indicates more low-frequency mutations than expected under a standard neutral model with no selection and a constant population size (Tajima, 1989). This genetic signature can be the result of a population expansion, purifying selection, or a very recent selective sweep. Conversely, *Tajima's D>0* indicates more intermediate- or high-frequency mutations than expected under a neutral model (**Figure 3.S3 p.56**). It can be explained by population contraction, balancing selection, or negative frequency-dependent selection.

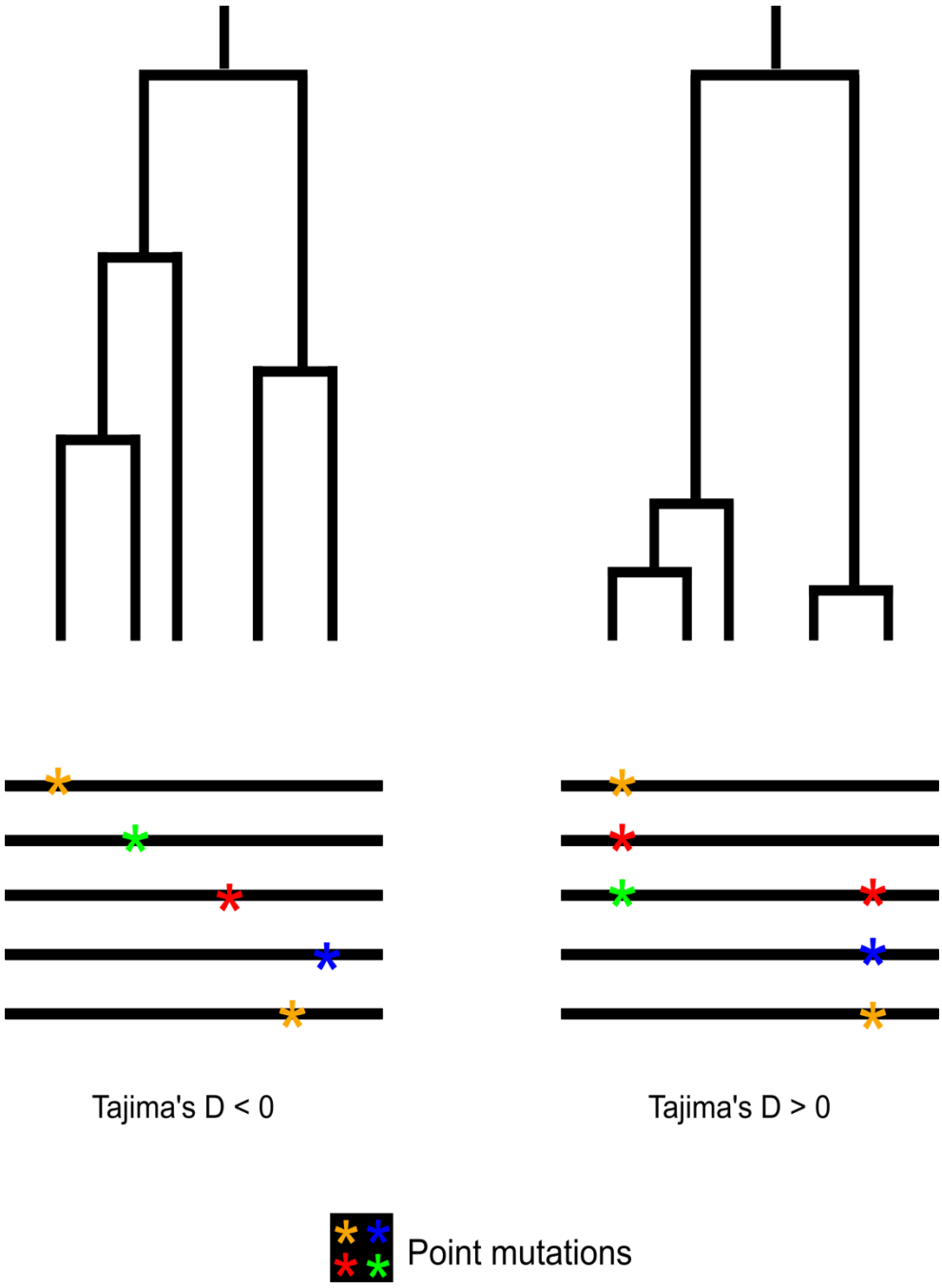


Figure 3.S3 Underlying logic of *Tajima's D*

Tajima's D measures the difference between average per-site pairwise differences of gene alleles (θ_π) and the normalized number of polymorphic sites (θ_w). It takes negative values when there are more low-frequency mutations than expected under a standard neutral model with no selection and a constant population size (Tajima, 1989). In a phylogenetic tree of the gene alleles, this would result in high divergence and long branches as the low-frequency mutation are rarely shared across

alleles and occurs at different polymorphic sites most of the time, which can be observed in the corresponding alignment. This genetic signature can be the result of a population expansion or a very recent selective sweep. Positive *Tajima's D* values are observed when there are more intermediate/high frequency mutations than expected under a neutral model (Tajima, 1989). In a phylogenetic tree of the gene alleles, this would result in low divergence and shorter branches as the intermediate/high frequency mutations are frequently shared across alleles and occurs at few polymorphic sites, which can be observed in the corresponding alignment. Positive values of *Tajima's* D can be explained by population contraction or negative frequency-dependent selection.

The above metrics were calculated for every gene in each sample by mapping metagenomic reads and calling SNVs after applying a 10X sequencing coverage filter (Methods). Consistent with previous estimates across multiple kingdoms of life (Koonin & Wolf, 2010), we observe that θ_π and θ_w distributions across samples span 3 to 4 orders of magnitude (**Figure 3.S4 p.58**). Also consistent with previous estimates in bacteria over different time scales (Gardon et al., 2020; Garud & Pollard, 2020; Sela et al., 2016), *dN/dS* tends to be less than one, suggesting the predominance of purifying selection at the protein level (**Figure 3.S4 p.58**). Our estimates of these population genetic metrics from metagenomic data are thus within an expected range and appear to behave as expected.

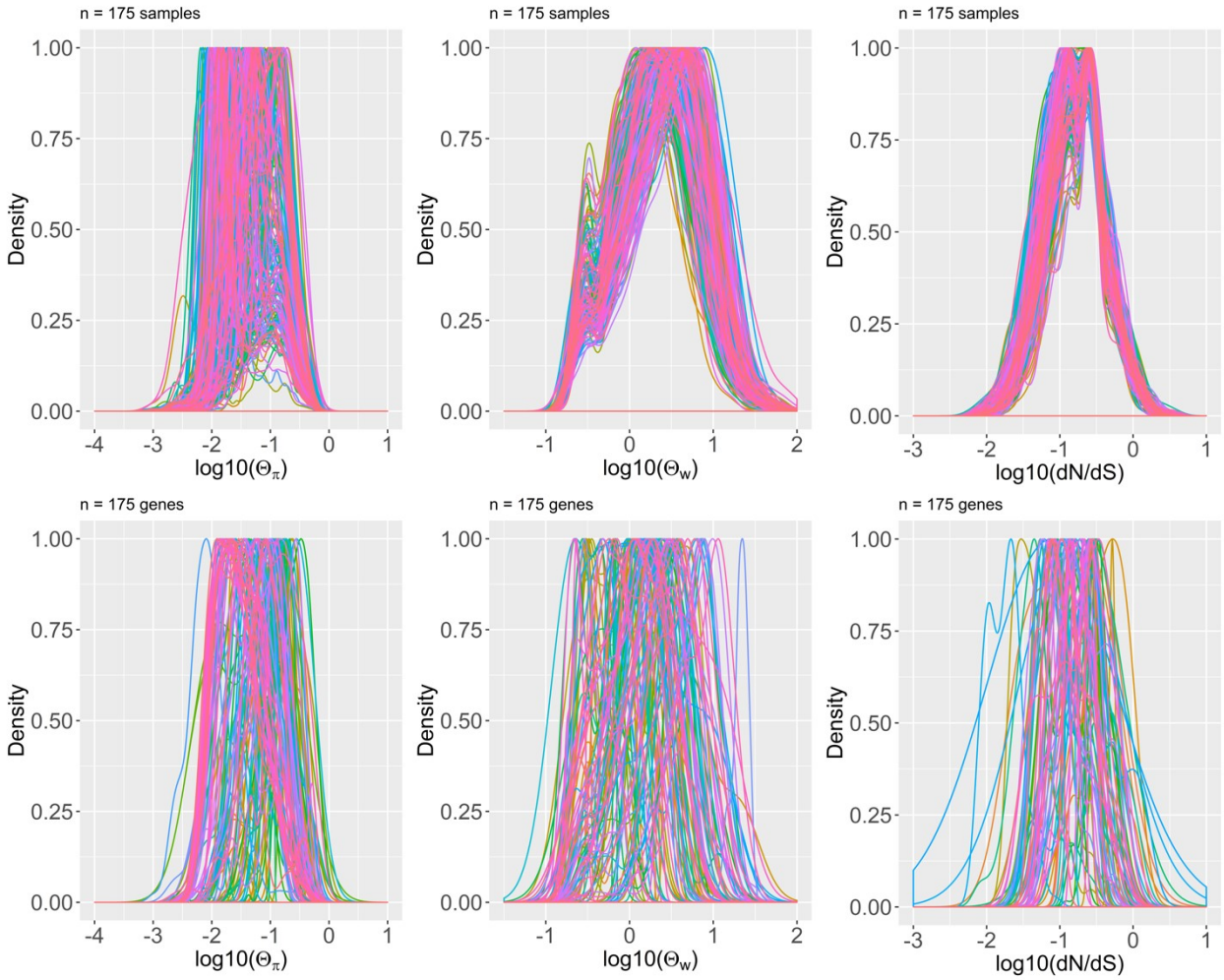


Figure 3.S4 Population genetics metrics distributions vary more across genes than across samples (people)

We measured and compared the variation of mobile genes population genetics metrics across samples/individuals and across genes. The top row represents the distributions of *dN/dS*, θ_π and θ_w (logarithmic scale) across samples and the bottom rows represents the distribution of these descriptors across genes (downsampled to 175 to match the sample size). Each column represents one of three replicates random downsamplings. It can be visually seen that the distributions are more variable across genes (bottom row) than individuals (top row), and this is quantified by the Kolmogorov-Smirnov *D* statistic, computed for each of 999 replicate downsamplings to produce **Figures 3.2 (p.59)**.

3.3.3 Population genetic metrics vary more across mobile genes than across host attributes

With these metrics in hand, we asked whether mobile gene evolution is mainly driven by bacterial- or human host-specific selective pressures. To do so, we determined whether population genetic metrics varied more across gene families or across individuals. We first compared distributions of pairwise differences for each metric using the Kolmogorov-Smirnov test, and found much greater variation between genes than between individuals (**Figures 3.2 p.59 and 3.S4 p.58**). This result indicates that, on short time scales, the selective pressures quantified by the four metrics may be less affected by person-specific factors, such as lifestyle or social networks, than by gene functions within a microbial cell. In other words, although some mobile genes may enable adaptations to personalized factors such as diet (Brito et al., 2016), sequence evolution is relatively unaffected by these factors on short time scales (within an individual). In contrast, population genetic metrics vary substantially more across genes, suggesting that selective pressures act predominantly at the level of gene function.

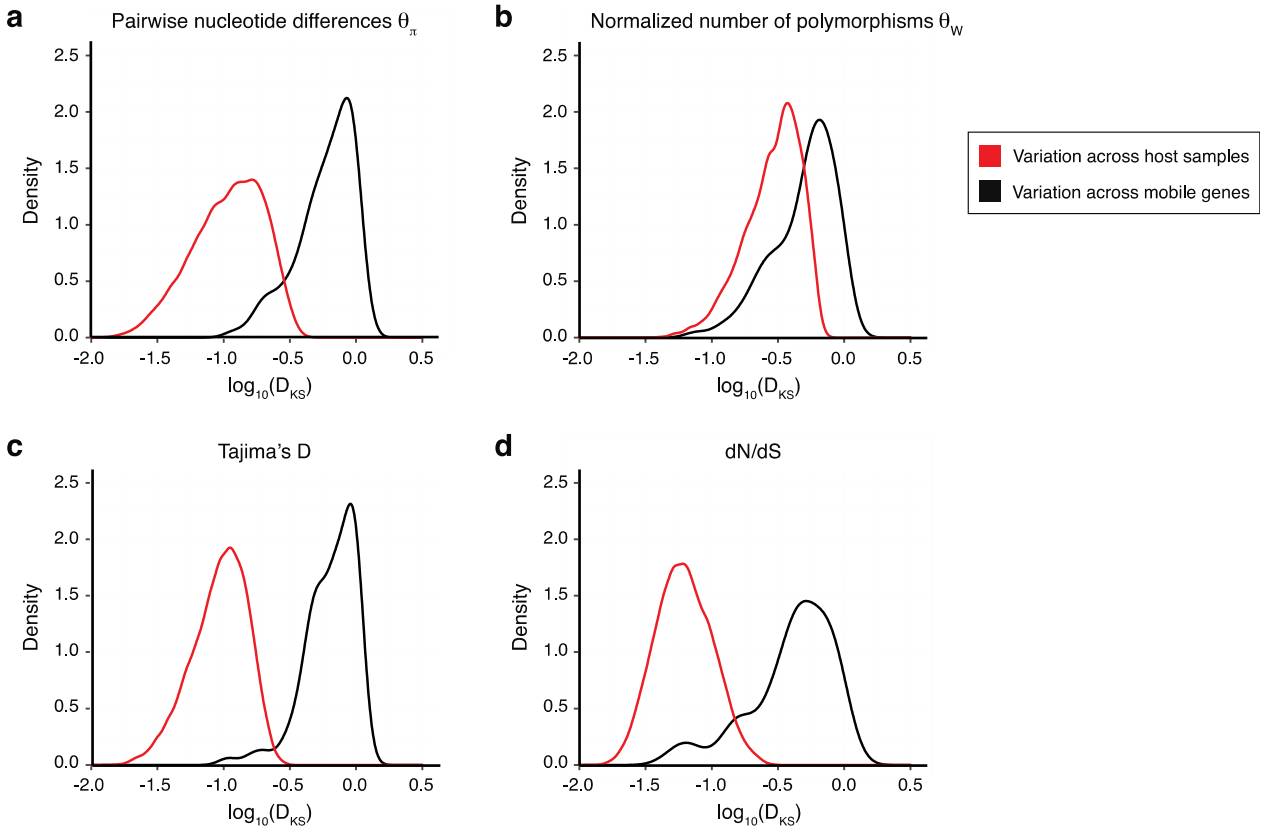


Figure 3.2 Mobile gene evolution varies more widely across genes than across samples (people)

Each panel shows the distribution of the variation of population genetic metrics among samples (red) or among gene families (black) through the distribution of $\log_{10}(D_{KS})$ statistics. The D_{KS} statistic from the Kolmogorov-Smirnov test measures the maximal distance between a pair of cumulative distributions – in this case, across either samples or genes. Panels a, b, c and d represent the variation of θ_π, θ_w, *Tajima's D* and *dN/dS* respectively. We downsampled the 37 853 genes to the same size as the number of samples set to avoid the potential bias toward more variation in the larger dataset of genes (999 sub-samples). This figure presents the result for 999 sub-samples of 175 genes and shows that there is more variation across genes than across samples/individuals for all the population genetics metrics (KS test, $P < 2.2 \times 10^{-16}$). See **Figure 3.S4 (p.58)** for example distributions across genes and samples.

To validate that person-specific factors have weak effects on mobile gene sequence evolution, we used a linear regression where the continuous response variable is one of the population genetics metrics and the qualitative/categorical explanatory variable is a host attribute (Methods). Because the statistical significance of such an analysis is affected by sample size, we selected mobile genes with less than 30% missing values across the 172 samples for which metadata were available, for a total of 1333 tested genes. Host age and sex did not show any significant effects on mobile gene sequence evolution. However, a person's household or village significantly influenced the evolution of just a few mobile genes (1.13% to 4.25% of the 1333 tested genes; **Figure 3.3A p.61**). In this small subset of significant genes, the correlations between population genetic metrics and household (adjusted R^2 ~0.6 to ~0.68) were stronger than correlations with village (adjusted $R^2 < 0.3$), and these results were robust to varying the quality filters applied to the data (**Figure 3.S5 p.63**). The small set of genes significantly influenced by household and village could be representative of very specific family/village selective pressures such as diet. Annotations of these genes show that they are involved in a set of functions involved in carbohydrates, lipids, secondary metabolites and ions transport or metabolism, and potential antibiotic resistance through ABC-type multidrug transporter system (**Tables 3.S2 p.119**). Some of these functions are similar to those identified by Brito *et al.* as differentially abundant among villages or households [1]. Therefore, although village- or household-specific selective pressures do not explain much of the variation in population genetic metrics across genes, we cannot exclude

rare instances in which social networks or lifestyles drive the evolution of few mobile genes over short time scales.

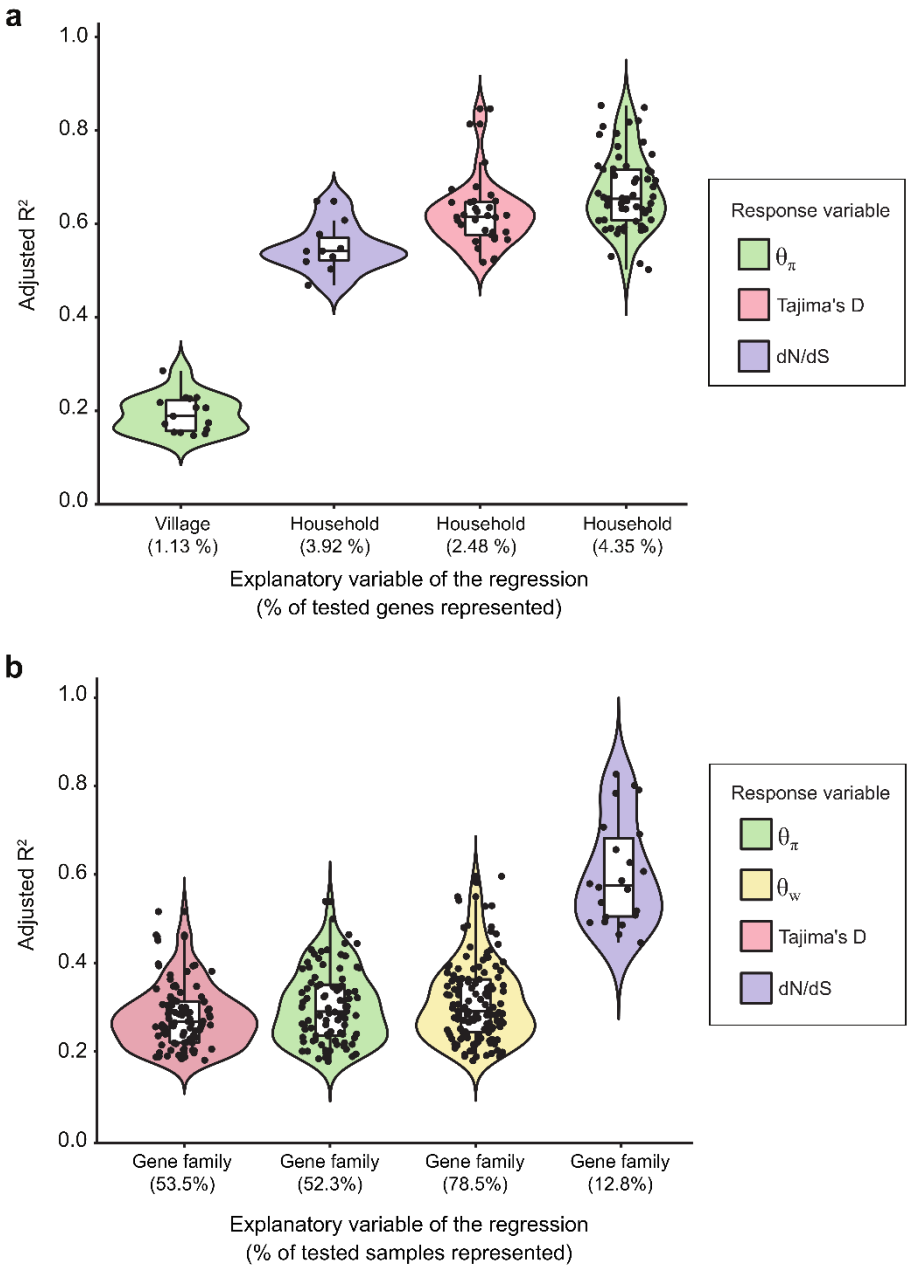


Figure 3.3 Gene function explains more variation in mobile gene sequence evolution than host attributes

A) Adjusted R^2 values for the categorical regressions between population genetic metrics (color-coded) and host attributes. We only considered genes with at least 10X coverage in a sample, and

we also required that mobile gene should have less than 30% missing values across samples, for a total of 1333 genes included in this analysis. The four strongest and most prevalent correlations between population genetics metrics and host factors are shown. Not shown are village significantly correlated with θ_w (0.15% of genes), *Tajima's D* (0.15%) and *dN/dS* (0%) and household significantly correlated with θ_w (0.23%). Host age and sex did not show any significant effects on mobile gene sequence evolution. Each black point represents a mobile gene for which the categorical regression is significant. The percentage of significant genes out of the total number of genes tested is indicated in parentheses along the x-axis. For *dN/dS*, the sample size was reduced to n = 255 genes because an additional filter requiring mutations to be seen in a least 5 metagenomic reads was applied before computing *dN/dS*, which can other be sensitive to sequencing errors (Methods)**.** B) Adjusted R^2 values of the categorical regressions between a population genetic metric and the gene family. Each black point represents a sample for which the categorical regression is significant. The percentage of significant samples out of the total number of samples tested is indicated in parenthesis along the x-axis. Only 172 out of 175 samples for which metadata was available are included in this analysis. We only considered genes with at least 10X coverage in a sample. We only included genes with a gene family annotation and required that each gene family be represented by at least 2 genes. Finally, we only included genes present in 30% or more of the samples, for a total of 512 genes included in the analysis.

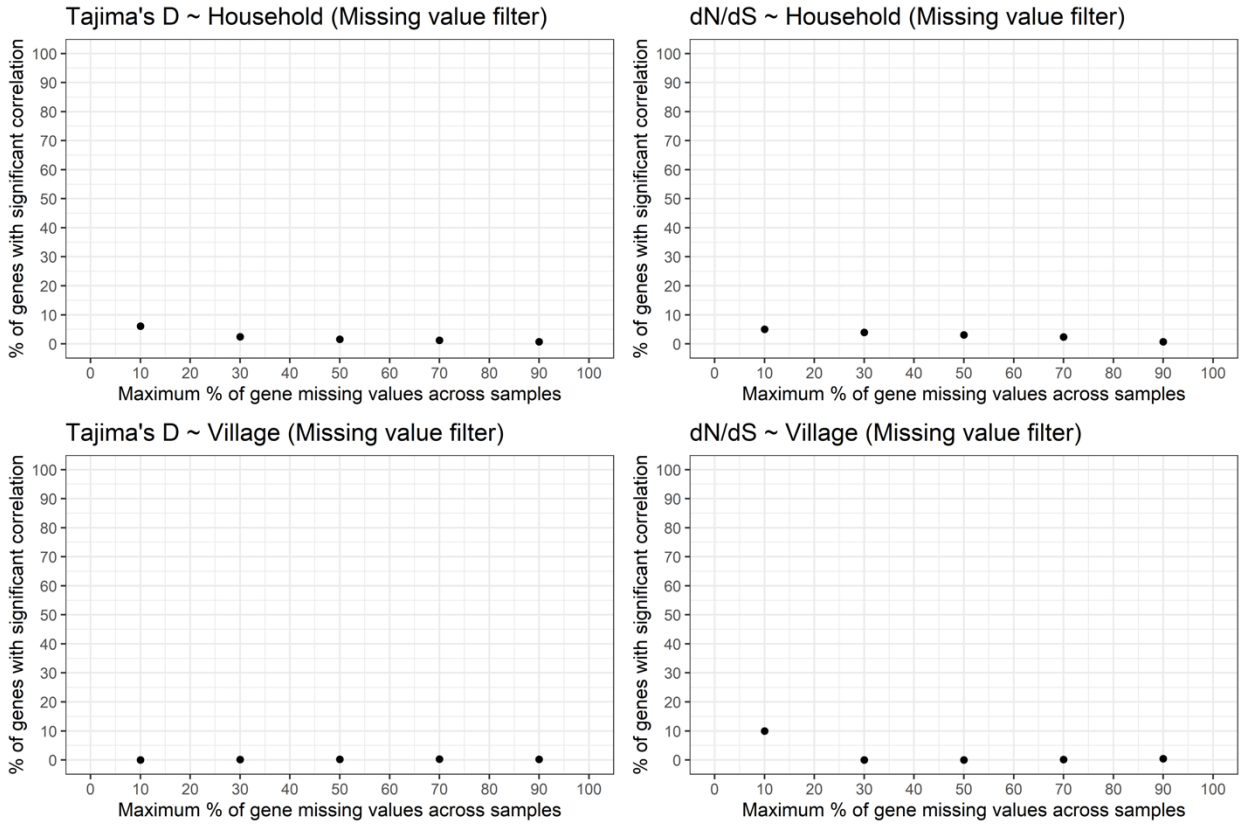


Figure 3.S5 The lack of significant correlations between host factors and mobile gene evolution is robust to filters imposed on missing data

This figure illustrates the percentage of genes for which the correlations between population genetics metrics and host factors are significant depending on the missing value filter stringency. The missing value filter defines a maximum percentage of samples in which the gene is absent or sequenced with less than 10X coverage. Increasing this threshold should increase the number of tested genes. **Figure 3.3A (p.61)** shows that, using a 30% missing value threshold, population genetics metrics do not significantly correlate with host factors, except for a minority of genes (<5% of tested genes). This figure shows that the non-significance of this correlation is robust to the missing value filters because no matter the threshold chosen, the proportion of genes for which the correlations is significant is still very small. Because *Tajima's D* captures information from both θ_π and θ_w, their correlations with host factors were not included in this figure.

Although host factors seem to have relatively little effect on the sequence evolution of most mobile genes on short time scales, selective pressures at the level of the genes might be more

important. Indeed, we observed higher variations of population genetics metrics between genes than between samples (**Figures 3.2 p.59**), which could be explained by gene attributes such as their cellular function. To test this hypothesis, we used linear regressions between population genetics metrics and gene families based on the following set of conditions:

(1) the gene should have at least 10X coverage to limit the impact of sequencing errors and to have confidence in the variant calling,

(2) the gene should have an available gene family annotation, which is the explanatory variable of the regression. Gene families annotations come from COG, KEGG, TIGRFAM , PFAM or dbCAN databases (Brito et al., 2016),

(3) the gene family should be represented by at least 2 genes within the dataset to avoid low sample sizes, and,

(4) the mobile gene should have less than 30% missing values across samples, for a total of 512 tested genes.

In contrast to human factors, gene functions defined by gene families from COG, explained more of the variation in mobile gene sequence evolution across samples. For θ_w, θ_π, and *Tajima's D*, gene families explained from ~20% to ~60% of the variance in >50% of the samples (**Figure 3.3B p.61**). For *dN/dS*, gene families explained up to 83% of the variance in 12.8% of samples. To ensure that this result was robust to differential sampling of genes (n=512) and individuals (n=172) in this analysis, we downsampled to n=172 genes and confirmed that human host factors explain much less variation in mobile gene evolution compared to gene functions (**Figure 3.S6 p.65**). A caveat of this analysis is that the explanatory power and the reproducibility of the correlation, i.e. the number of samples for which the correlation is significant (Bonferroni-corrected p-value < 0.05), depends on the balance between sample size (number of selected genes) and the stringency of the filters (**Figure 3.S7 p.67**). Indeed, as the filters stringency increases, more prevalent mobile genes or gene families that are better represented in the dataset are selected for the analysis and the correlation R^2 tends to increase. However, the sample size and the reproducibility of the correlation tend to decrease (**Figure 3.S7 p.67**). For instance, as the stringency of the missing value filter increases, fewer samples show significant correlations between *Tajima's D* and COG function, going from 88.4% of samples when a gene can be absent in at most 30% of samples (**Figure 3.3B**

p.61) to 40.1% significant when a gene can be absent in at most 10% of samples (**Figure 3.S7 p.67**). Although the correlation strength depends on the filters stringency, the median adjusted R^2 of the correlation is always higher than 20% and can reach up to ~60%. Altogether, these results suggest that COG functions appear to explain much of the short-term molecular evolution of a subset of mobile genes.

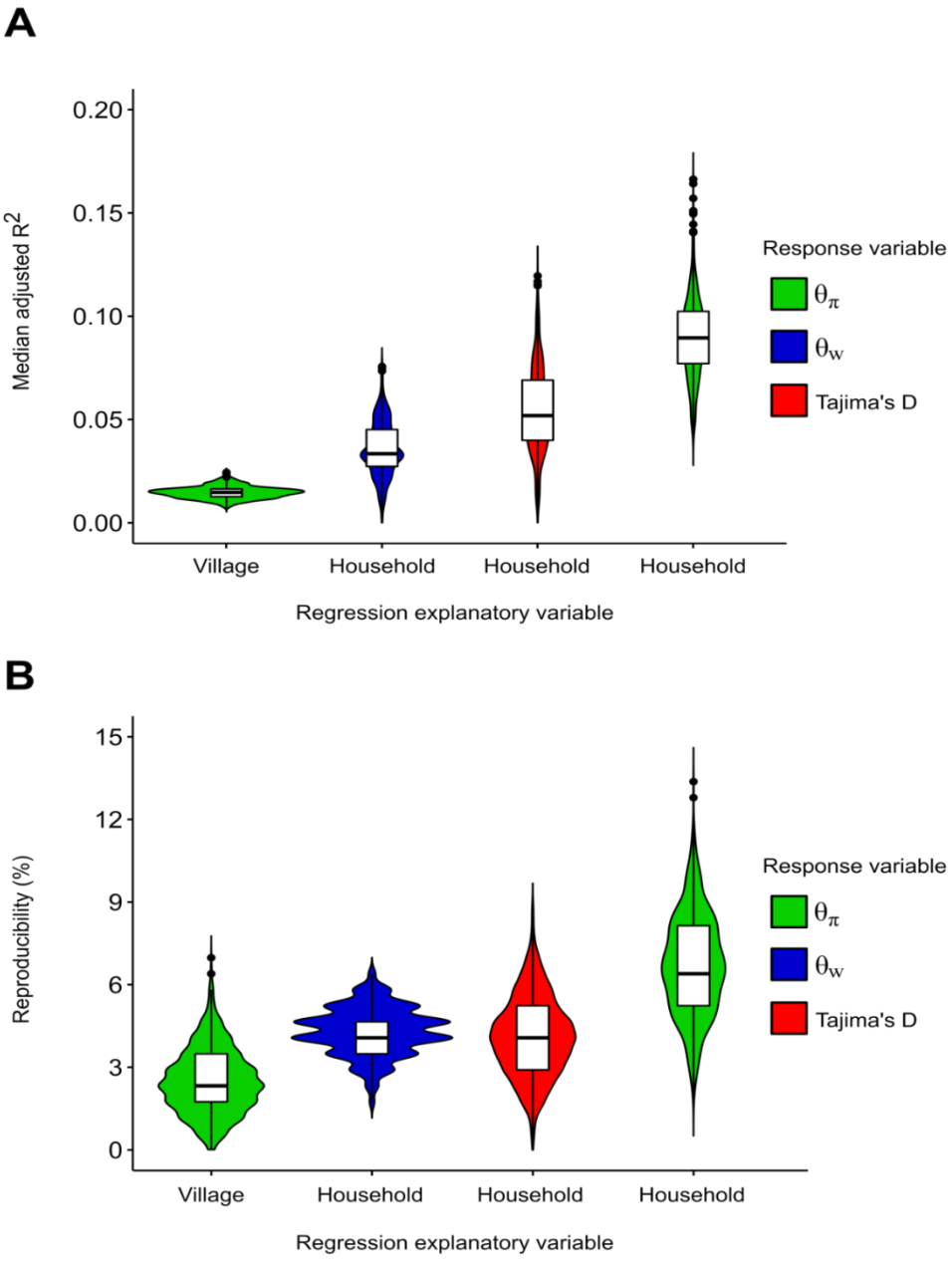


Figure 3.S6 Gene set size bias does not explain host attributes weak influence on mobile gene sequences evolution

We hereby want to make sure that host attributes correlations with population genetics metrics were not weaker than gene family correlations only because their significance was assessed on a bigger set of objects (1333 genes vs 172 samples). We evaluated the significance of these correlations with the adjusted R^2 and the reproducibility, i.e. the proportion of regression objects (genes or samples) for which the regression is significant. Therefore, there could be a bias toward observing lower reproducibility in bigger set of objects. A) Median adjusted R^2 across subsamples. This figure shows the mean adjusted R^2 of top 4 regressions between population genetics metrics (color-coded) and host attributes across the 999 subsamples (n=172 genes). The downsampling do not increase the regression significance and rather supports the fact that it is weak because the median adjusted R^2 are low (<= 0.2). B) Reproducibility (%) across subsamples. Reproducibility represents the percentage of genes for which host attributes correlations are significant. The correlations and genes represented are the same than in **Figure 3.S6A (p.65)**. This figure shows that the reproducibility of these correlations, which ranges between 0 and 15%, is lower than the reproducibility of gene family correlations (**Figure 3.3B p.61**), even after downsampling.

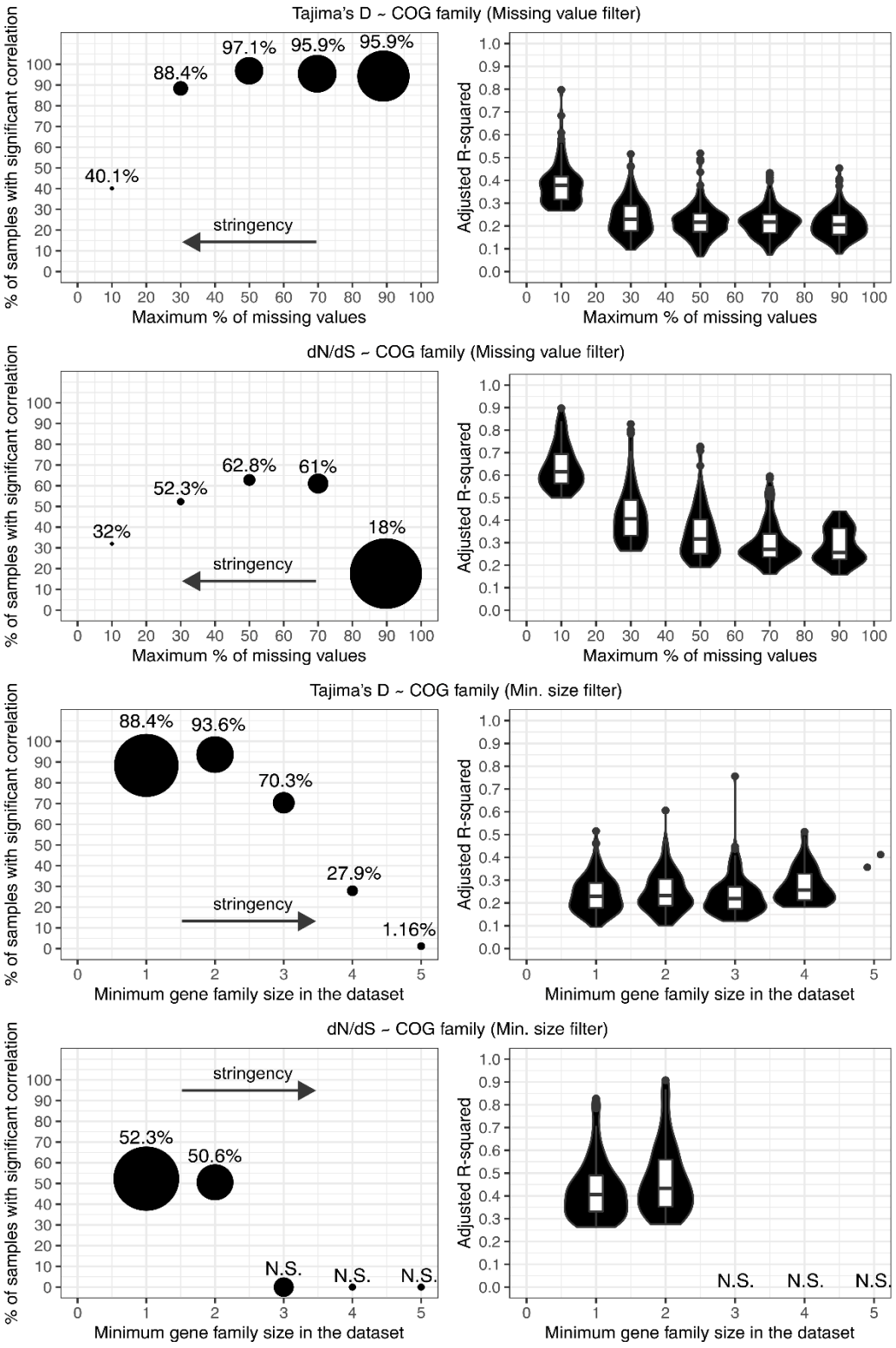


Figure 3.S7 The measured impact of gene family on mobile genes evolution depends on a trade-off between sample size and filters stringency

This figure illustrates the reproducibility, i.e. the % of samples with a significant correlation (FDR-corrected p-value < 0.05) , and the strength of the correlations between population genetics metrics and gene family depending on the missing value filter (1st and 2nd rows of the figure) and a filter on the minimum number of genes representing a gene family in the dataset (3rd and 4th rows of the figure). The size of the points represents the relative sample size used for the correlations (normalized number of genes that passed the filters) and the Y-axis value is indicated above each point for clarification. The missing value filter defines a maximum percentage of samples in which the gene is absent or sequenced with less than 10X coverage. Thus, increasing this threshold makes this filter decrease in stringency and thus increase sample size. As for the filter on the minimum size of a gene family within the dataset, it defines a minimum threshold such that increasing it would increase stringency and thus decrease sample size. These filters have been chosen respectively to handle missing values caused by gene absence across sample or gene with low coverage in gut metagenomes and to avoid the random effects of small sample size for gene families that are underrepresented in the dataset. We tested the influence of these filters one at a time while keeping the other parameter constant at its original value (<=30% missing value across samples for each selected gene and gene families with at least 2 genes in the dataset). The values of the parameters that yield no significant correlation are labeled as "N.S.".

3.3.4 Higher gene mobility is associated with low-frequency SNVs in the gut microbiome

In addition to gene- or environment-specific selective pressures, the rate of HGT is also expected to affect mobile gene molecular evolution, as it allows genes to spread across different species, possibly altering their population size and thus the efficacy of selection (Shapiro, 2017; Vos et al., 2015). To first order, each human host represents a distinct short-term evolutionary trial. Thus, to study the influence of HGT rate on molecular evolution within each of the human guts sampled, we correlated gene mobility with the population genetic metrics described above: *dN/dS*, θ_π, θ_w, and *Tajima's D.* All correlation results reported below are robust whether or not we include gene length and coverage as covariates in linear regressions (**Figure 3.S8 p.69**).

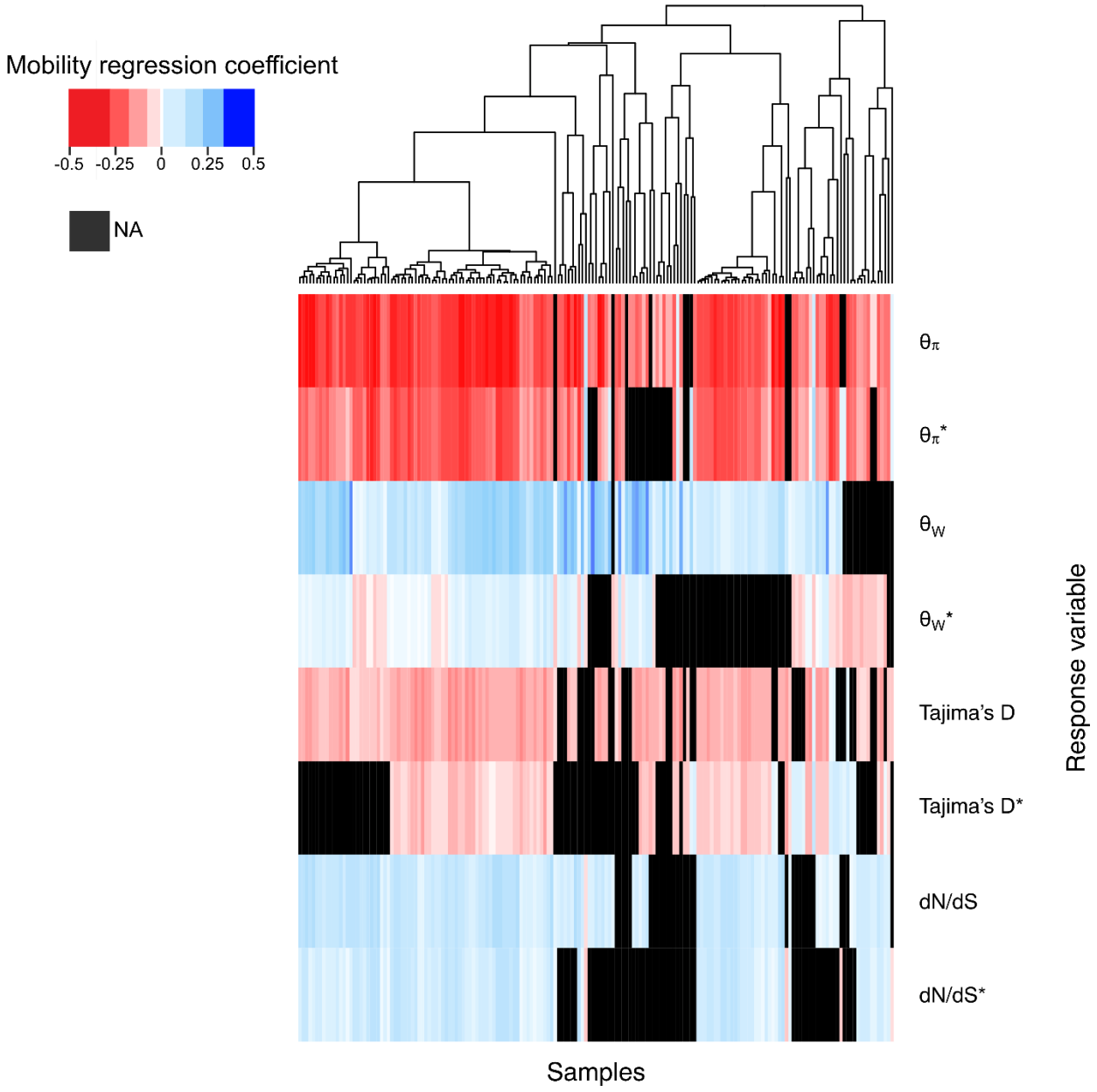


Figure 3.S8 Complete heatmap of gene mobility regression coefficients

For each response variable Y tested (θ_π, θ_w, *dN, dS, dN/dS* and *Tajima's D*), two regression models are performed: "Y ~ Gene Mobility" and "Y^* ~ Gene Mobility + Coverage + Gene length". The second type of model allows us to see the correlation between gene mobility and Y by controlling for coverage and gene length and thus sequencing artefacts like sequencing errors that increase with coverage and gene length. The heatmap contains non-significant regressions results

after FDR correction for multiple testing (black), negative significant correlations (red) and positive significant correlations (blue).

Using this regression approach, we first observed that the correlation between *dN/dS* and gene mobility was significant and positive in 144 out of 175 samples (**Figure 3.S8 p.69**), but with a low average adjusted R^2 of 0.03 (s.d. = 0.02). This correlation could be explained by the fixation of slightly deleterious non-synonymous mutations in the early stage of a population expansion (Parsch, Zhang, & Baines, 2009) as it is the case when mobile genes are spreading across species on short time scales. Alternatively, this result could also be explained by slightly increasing positive or relaxed purifying selection with increasing gene mobility, but we refrain from drawing strong conclusions due to the weak R^2 values. We next observed that 159 out of 175 samples had a somewhat stronger significant correlation between θ_w and gene mobility (linear regression with Bonferroni-adjusted p-value < 0.05), and all the significant correlations were positive (mean adjusted R^2 = 0.06; s.d. = 0.06). This is consistent with a model in which mobile genes accumulate SNVs that remain at low frequency (as measured by θ_w, which is sensitive to these low-frequency mutations) as they spread across species. We also observed that θ_π, which is more sensitive to intermediate-frequency mutations, decreases with gene mobility (**Figure 3.4A p.71**). Among samples in which θ_π versus gene mobility regression results were significant (164 out of 175 samples with Bonferroni-adjusted p-value < 0.05), ~95% of them exhibited this negative correlation (mean adjusted R^2 = 0.08; sd = 0.05). As a result, *Tajima's D* (which reflects the difference between θ_π and θ_w) is significantly negatively correlated with gene mobility in ~83% of samples (**Figure 3.4A p.71**). Even if the R^2 value are modest, we note that the trends are highly repeatable across samples. Reasons for the relatively low R^2 values could include noise in the gene mobility metric (based on a small sample of genomes) and/or variable selective pressures across genes. There are several reasons for this enrichment of low-frequency SNVs (resulting in lower *Tajima's D* values) in more mobile genes, including purifying selection keeping deleterious mutations at low frequency, recovery of new polymorphism after a recent selective sweep, or population expansion. This result suggests that HGT can spread genes across species faster than SNVs are able to rise to high frequency.

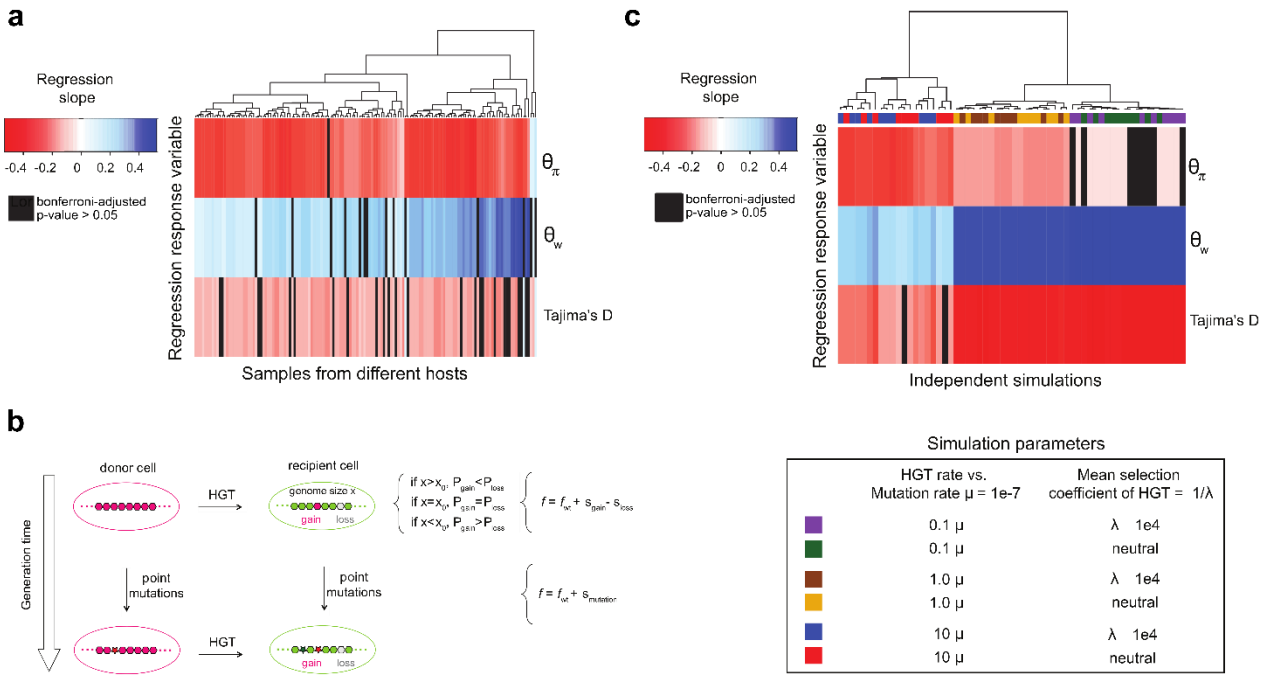


Figure 3.4 Gene mobility is negatively correlated with *Tajima's D* in real and simulated microbiomes

A) Real data from Fiji. The heatmap shows the slope of a regression model in which either θ_π, θ_w or *Tajima's D* is the response variable and gene mobility is the explanatory variable (across samples). Regression p-values were obtained through a *t*-test. The heatmap contains non-significant regressions results after Bonferroni p-value filter (black), negative significant correlations (red) and positive significant correlations (blue). Data standardization was performed before each regression to respect the *t*-test's assumption of normality and we validated that it converges with a non-parametric test (Methods). Heatmap rows and columns were clustered with Euclidean distance and complete linkage clustering.

B) Representation of simulation events over two generations. In the first generation, a gene gain event occurs through HGT. Gene gain is represented by the transfer of gene from a donor cell to a recipient cell and increases the genome size of this recipient cell. The probability of future gene gain or gene loss events (P_{gain} and P_{loss} respectively) is determined by the difference between the current genome size of the cell (x) and the equilibrium genome size (x_0). At equilibrium, the probability of gene gain and gene loss is the same by definition ($P_{gain}=P_{loss}$). An increase of genome size until it exceeds the equilibrium point ($x > x_0$) leads to gene loss being more likely than gene gain ($P_{gain}<P_{loss}$). Gene gain also increases the fitness ($f > f_{WT}$) of the recipient cell based on the

selection coefficient of the transferred gene (s_{gain}). In the model, each gene has its own selective coefficient which is drawn from an exponential distribution $exp(\lambda)$ with an expected value of $1/\lambda$. Gene gain is either slightly beneficial or neutral in this model and has the opposite fitness effect of gene loss, which is slightly deleterious or neutral ($-s_{gain} = s_{loss}$ where $s_{gain} >= 0$). Gene loss decreases the genome size of the target cell and in case this decrease leads to a smaller genome size than equilibrium, the probability of gene gain becomes higher than the probability of gene loss ($P_{gain} > P_{loss}$). Gene loss also decreases the fitness of the target cell ($f < f_{WT}$) based on the selection coefficient of the lost gene (s_{loss}). Finally, as represented in the second generation, mutations can also occur and change the fitness of the cell based on a selective coefficient ($s_{mutation}$) which is drawn from a distribution (Methods).

C) Simulated data. The heatmap shows the slope of a regression model in which either θ_π, θ_w or *Tajima's D* is the response variable and gene mobility is the explanatory variable (across simulation replicates). Simulations with different parameter for HGT rate and or distributions of selective coefficients ($s \sim exp(\lambda)$) are color-coded (n=10 replicates per simulation).

3.3.5 A subset of gene functions experiences a divergent regime of natural selection

Having established that *Tajima's D* correlates negatively with gene mobility while coverage tends to correlate positively with mobility (**Figure 3.1 p.52**), we sought to determine if these general trends apply equally to all gene families. While the trends are significant across samples, the large variations observed across genes (**Figure 3.2 p.59**; **Figure 3.S4 p.58**) could represent evolutionary regimes that are specific to some gene families. To test this hypothesis, we used linear mixed models with gene mobility as a predictor of either *Tajima's D* or coverage as a response variable, while controlling for random variations across gut microbiome samples and allowing the response to vary across COG categories (Methods). This analysis was performed on genes with at least 10X coverage and available COG annotations (n= 3608 mobile genes).

As expected, based on the overall positive relationship observed (**Figure 3.1 p.52**), coverage and gene mobility are positively and significantly correlated across most COG categories

(**Figure 3.5A p.74**). COG category X (mobilome, prophages, and transposons) stood out as the strongest contributor to this positive relationship, consistent with this signal being driven by genes with the highest mobility. Removing sample identity or COG category from the linear mixed models significantly decreased the fit of the models, suggesting that they both significantly contribute to explaining variation in the mobility-coverage and Tajima's D-coverage relationships (**Tables 3.S3A p.75 and 3.S3B p.76**). We also confirmed that *Tajima's D* is negatively correlated with gene mobility (**Figure 3.5A p.74**), as observed in the regression analysis (**Figure 3.4A p.71**). Deviations from this correlation could thus reveal signatures of selection that are specific to certain gene families. These COG categories for which *Tajima's D* significantly increases with mobility, include P (Inorganic ion transport and metabolism), I (Lipid transport and metabolism), Q (Secondary metabolites biosynthesis, transport and catabolism), V (Defense mechanisms) and O (Posttranslational modification, protein turnover, chaperones), representing ~30% of gene families (**Figure 3.S9 p.77**). There are several explanations for why these gene families maintain or accumulate intermediate-frequency SNVs (*i.e.* an increase in *Tajima's D*) while being transferred to many new species (**Figure 3.5B p.74)**. The first explanation is a population contraction, or in this context, a reduction of the number of gene copies across species. However, this is unlikely for these subsets of genes because their coverage, which is a proxy of the relative abundance, increases with mobility. The second explanation is that these genes could be subject to species-specific selective pressures that fix mutations in some species but not others, resulting in intermediate SNV frequencies in the bulk metagenome. The third potential explanation is that negative frequency-dependent selection, which is thought to be an important force shaping pangenome evolution (Cordero & Polz, 2014; Domingo-Sananes & McInerney, 2019), is acting on these genes, within species, between species, or both. Thus, the last two scenarios, which rely on the presence of distinct selective pressures on these subsets of genes, most likely explain how some mobile genes can maintain or accumulate intermediate-frequency SNVs as they spread across species.

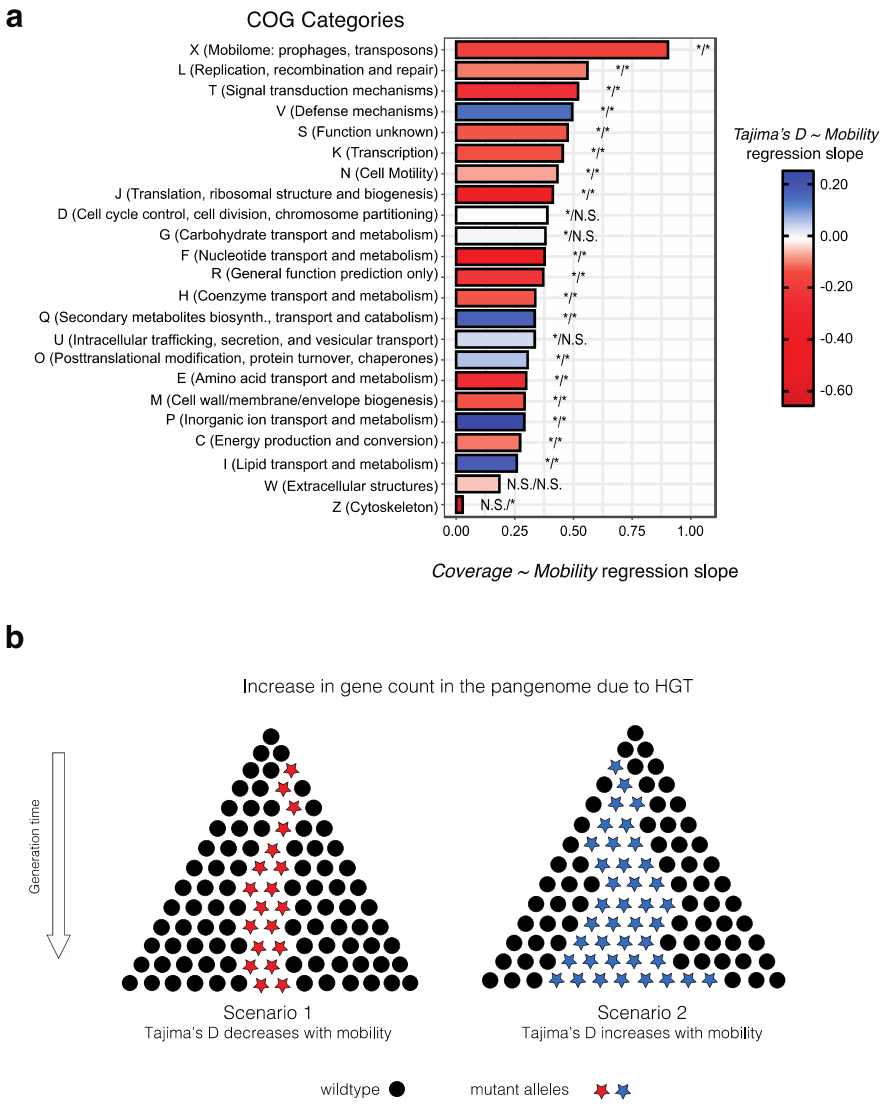


Figure 3.5 Gene mobility regressions reveal a minority of genes with distinct signals of selection

A) Linear mixed model regression slopes per COG category**.** This figure illustrate COG categories regression slopes for the linear mixed models $Coverage \sim Gene\ mobility + Sample + COG\ category$ and $Tajima's\ D \sim Gene\ mobility + Sample + COG\ category$ with $Sample$ and $COG\ category$ being considered as random effects. Data were normalized using the Box-Cox transformation to ensure the condition of residual normality was accounted for before building the linear mixed model (Coverage Box-Cox λ = -0.01; Gene mobility Box-Cox λ = -0.005). We only used the 99.6% of Tajima's D values that were negative and thus inversed their sign before applying Box-Cox transformation, which only works with positive values. We then performed the linear

mixed model regression $-Tajima's\ D \sim Gene\ mobility + Sample + COG\ category$ and inversed the sign of its slope (-*Tajima's D* Box-Cox λ = 2). The sign of the slopes was consistent with simple linear regressions. The asterisks at the tip of each bar indicate the significance of the simple linear regressions $Coverage \sim Gene\ mobility$ and $Tajima's\ D \sim Gene\ mobility$ respectively for the associated COG category (*=Significant; N.S. = Not Significant; Cut-off: Bonferroni-adjusted p-value < 0.05).

B) Schematic of the evolutionary scenarios compared using linear regressions. Scenario 1 represents the situation in which mobile genes *Tajima's D* is negatively correlated with gene mobility because HGT is faster than fixation of mutated alleles (red stars). Scenario 2 represents the situation in which *Tajima's D* correlates positively with mobility. These genes maintain intermediate frequency mutations (blue stars) despite being frequently transferred to new species due to negative frequency-dependent selection or species-specific selective pressures that fix mutations in some species but not others. Note that the gene copies (dots or stars) illustrated here could come from members of the same or different species in the microbiome.

Model	Description	R^2	Log-likelihood	LRT p-value (Nested model vs M1 model)
M1	*Tajima's D* vs Mobility + all random factors, i.e. Sample and COG category	0.07	-454 557	-
M2	*Tajima's D* vs Mobility + COG category (without Sample)	0.04	-457 645	$< 2.2x10^{-16}$
M3	*Tajima's D* vs Mobility + Sample (without COG category)	0.033	-458 084	$< 2.2x10^{-16}$

Table 3.S3A Regression strength of the linear mixed model $Tajima's\ D \sim Mobility + Sample + COG\ category$ and nested models LRT

Because the likelihood ratio test (LRT) p-value is highly significant for each of the nested models, all the random effects tested, i.e. sample and COG category, have a significant effect on the regression. M2 model R^2 is much higher than M3 model R^2, suggesting that the relationship between *Tajima's D* and Mobility varies more across COG categories than across samples. The R^2

of this linear mixed model, i.e. $Tajima's\ D \sim Mobility + Sample + COG\ category$, is much lower than the R^2 of the mixed model $Coverage \sim Mobility + Sample + COG\ category$ (**Table 3.S1 p.54**), because mobility correlates negatively with *Tajima's D* for some COG categories and positively for others (**Figure 3.5A p.74**) so that the general trend is a weak, but significant negative correlation, contrarily to Coverage which always correlates positively with mobility (**Figure 3.1 p.52**).

Model	Description	R^2	Log-likelihood	LRT p-value (Nested model vs M1 model)
M1	FPKM vs Mobility + all random factors, i.e. Sample and COG category	0.321	-334 6029	-
M2	FPKM vs Mobility + COG category (without Sample)	0.205	-345 6259	< 2.2x10-16
M3	FPKM vs Mobility + Sample (without COG category)	0.284	-336 2567	< 2.2x10-16

Table 3.S3B Regression strength of the linear mixed model $FPKM \sim Mobility + Sample + COG\ category$ and nested models LRT

FPKM represents the number of fragments that mapped to a gene sequence normalized by the length of the gene in kilobase and the number of reads produced from the sample in million of mapped reads. It is a common metric of gene relative abundance and it is positively correlated with mobility (p-value = 9.17E-07 < 0.05). Here we show that all the random effects tested, i.e. sample and COG category, have a significant effect on the regression because the likelihood ratio test (LRT) p-value is highly significant for each of the nested models. The M3 model R^2 is much higher than the M2 model R^2, suggesting that the relationship between FPKM and Mobility varies more across samples than across COG categories, as it is the case for Coverage.

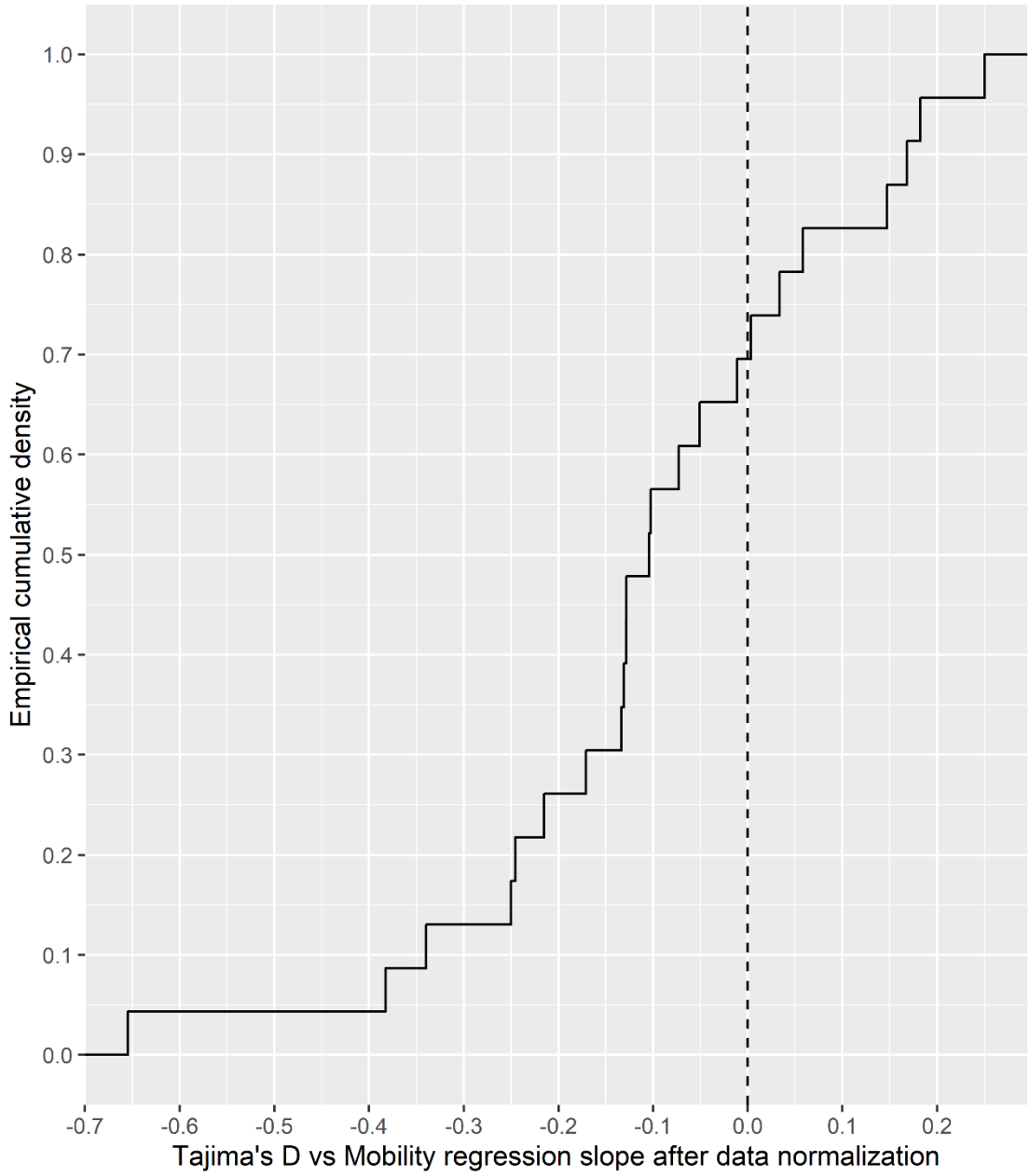


Figure 3.S9 Cumulative density distribution of $Tajima's\ D \sim Mobility$ regression slope across COG categories

3608 unique mobile genes data across the 175 samples were selected for this analysis. These mobile genes needed to pass the filters for variant analysis, i.e. mean site depth was >= 10, and have available COG annotations in the Metadata (Brito *et al.*, 2016). Data were normalized to make sure that regression residuals normality condition was respected before building the linear mixed model (Gene mobility Box-Cox λ = -0.005). We only used the 99.6% of Tajima's D values that were

negative and thus inversed their sign before applying Box-Cox transformation, which only works with positive values. We then performed the linear mixed model regression " −Tajima′s D ~ Gene mobility + Sample + COG category "and inversed the sign of its slope (-*Tajima's D* Box-Cox λ = 2). The sign of the slopes was consistent with simple linear regressions. COG categories, for which *Tajima's D* is positively correlated to mobility, which is a deviation from the general trend, represent 30% of the distribution. As mentioned in section 3, these genes are part of COG categories that are related to important beneficial functions in the gut microbiome, i.e. P (Inorganic ion transport and metabolism), I (Lipid transport and metabolism), V (Defense mechanisms), Q (Secondary metabolites biosynthesis, transport and catabolism), U(Intracellular trafficking, secretion, and vesicular transport), O(Posttranslational modification, protein turnover, chaperones) and G(Carbohydrates transport and metabolism).

3.3.6 Simple evolutionary simulations recapitulate the observed effects of HGT on mobile gene sequence evolution

To better understand potential mechanisms underlying the relationship between gene mobility and sequence evolution observed in the Fiji microbiome data, we implemented the explicit simulation of HGT and sequence evolution in SodaPop, a forward evolutionary simulation toolkit (Gauthier et al., 2019) (https://github.com/arnaud00013/SodaPop). Similar to Sela *et al.*, gene gain and loss are constrained to maintain genome size equilibrium and to have opposite fitness effects (**Figure 3.4B p.71**) (Sela et al., 2016). We used an updated version of the Sodapop model, which originally simulates protein sequence evolution with the distribution of fitness effects mutations derived from biophysics-based protein fitness landscapes (Gauthier et al., 2019). Briefly, we simulated a Wright-Fisher process for asexual populations (Gauthier et al., 2019) with 10 bacterial species. Each simulation included 5,000 cells in total, divided into 10 species, run for 10^5 generations. Each gene has an explicit sequence which evolves by a Jukes-Cantor point mutation model [28], including synonymous sites that do not affect fitness and nonsynonymous sites with a distribution of fitness effects of which 30% are lethal (Eyre-Walker & Keightley, 2007) (Methods). Genomes also experience HGT events, with explicit gene gain and loss events. The rates of these two events are updated at each generation for each cell to maintain an equilibrium around the genome size x_0, set to 500 genes (**Figure 3.4B p.71**) (Sela et al., 2016). Genomes larger than x_0 are prone to gene loss, but genomes smaller than x_0 are prone to gene gain. We also modeled gene gain and loss selection

coefficients, specific to each gene and drawn from an exponential distribution with parameter λ (Methods). We kept simulated population sizes small due to memory limitations. To make sure this limitation does not cause excessive effects of drift (*e.g.* the accumulation of deleterious mutations leading to extinction, also known as Muller's Ratchet (Bachtrog & Gordo, 2004)) we forced species relative abundances to remain constant. We also set a relatively high mutation rate of 10^{-7} mutations per site per generation to compensate for the small population sizes and to ensure that enough mutations were generated in a reasonable number of generations. Genome size equilibrium was reached for every simulation, indicating robustness to variable starting conditions (**Figures 3.S10 p.124**). Altogether, this model allows us to test if the relationships between gene mobility and population genetic metrics observed in the real data can be observed under varying rates of HGT and adaptive benefit of acquired genes.

We found that the simulation could recapitulate the major features observed in the real Fiji microbiome data without requiring that mobile genes provide adaptive value to a human host or to its bacterial genome. First, the simulations can recapitulate the shape of the observed distribution of gene mobility (**Figure 3.S1 p.51**). A caveat is that simulations are far from including all the complexity of the gut microbiome, *i.e.* the number of species, population structures and other features not simulated, and the distributions were only compared for one illustrative set of input parameters (**Figure 3.S1 p.51**). Thus, we do not claim that our model can provide a precise quantitative description of gene mobility in the gut microbiome, but rather that it can recapitulate the major qualitative features.

Second, the simulations recover the positive correlation between gene mobility and census population size (metagenomic coverage) observed in the real data (**Figure 3.1 p.52**). The positive correlation was always stronger in the simulations (mean adjusted R^2 of 0.705 across all parameter settings, standard deviation = 0.190) compared to the real data (mean adjusted R^2 of 0.085 across all parameter settings, standard deviation = 0.076). This suggests that factors not included in the model, such as negative frequency-dependent selection and noise in the gene mobility metric, reduced the strength of the correlation in the real data. The positive correlation was stronger in simulations with relatively lower HGT rate but was largely unaffected by whether HGT events were neutral or adaptive to host cell fitness (**Table 3.S4 p.80**). This suggests that relatively high

HGT rates could also explain the weaker correlation between gene mobility and coverage observed in the real data.

Simulation (n=10 replicates)	Average slope (standard deviation σ)	Average adjusted R^2 (standard deviation σ)	Average p-value (with Bonferroni correction)
HGT rate = 0.1μ; HGT λ = 1E4	0.958 (σ = 0.002)	0.917 (σ = 0.004)	$< 2.2x10^{-16}$
HGT rate = 0.1μ; neutral HGT	0.958 (σ = 0.002)	0.917 (σ = 0.004)	$< 2.2x10^{-16}$
HGT rate = 1μ; HGT λ = 1E4	0.855 (σ = 0.006)	0.731 (σ = 0.011)	$< 2.2x10^{-16}$
HGT rate = 1μ; neutral HGT	0.860 (σ = 0.006)	0.740 (σ = 0.010)	$< 2.2x10^{-16}$
HGT rate = 10μ; HGT λ = 1E4	0.684 (σ = 0.238)	0.467 (σ = 0.033)	$< 2.2x10^{-16}$
HGT rate = 10μ; neutral HGT	0.674 (σ = 0.251)	0.454 (σ = 0.034)	$< 2.2x10^{-16}$

Table 3.S4 Simulations support the positive correlation between gene census population size and Mobility

Gene census population size was estimated with the number of copies of the gene. Each simulation included 5000 cells, 10 species, 500 genes per cells at equilibrium and a simulation time of 10^5 generations. No matter if HGT was neutral or adaptive, this positive correlation was strongly supported (average adjusted R^2 over the 60 simulation replicates = 0.705; standard deviation = 0.190; p-value $< 2.2x10^{-16}$ for all replicates).

Third, we assessed whether the simulations could reproduce the observed correlations between population genetics metrics and gene mobility. Simulations recapitulated most of the

observed effects of HGT on nucleotide diversity in real data. Specifically, *Tajima's D* correlates negatively with gene mobility in simulations, with a median adjusted R^2 of 0.32 (mean = 0.23; sd = 0.13) compared to a median adjusted R^2 of 0.01 (mean = 0.01; sd = 0.01) in the real data and reproducible across ~87% of simulations compared to ~83% of the samples in the real data (**Figure 3.4 p.71**). The variation in this correlation is explained more by HGT rate than by HGT fitness effects (neutral or adaptive selective coefficients on gene gain/loss). This can be seen in the heatmap, in which simulations cluster by HGT rate rather than HGT fitness effect (**Figure 3.4C p.71**). Along the same lines, we performed a K-S test on the slopes of the regression between *Tajima's* D and mobility and observed that this slope varies more because of HGT rate than HGT fitness effect (**Figure 3.S11 p.82**). Simulations also predict that *dN/dS* also correlates positively but weakly with mobility, but only at intermediate HGT rates (**Figure 3.S12 p.83)**. A similar pattern is observed in the real data, in which *dN/dS* correlates weakly with mobility (**Figure 3.S8 p.69**). Overall, real microbiome data is recapitulated by our simple evolutionary model, which includes only selection for a stable genome size, without the need to invoke adaptive advantage of mobile genes to their bacterial genomes, or to include any human host factors whatsoever.

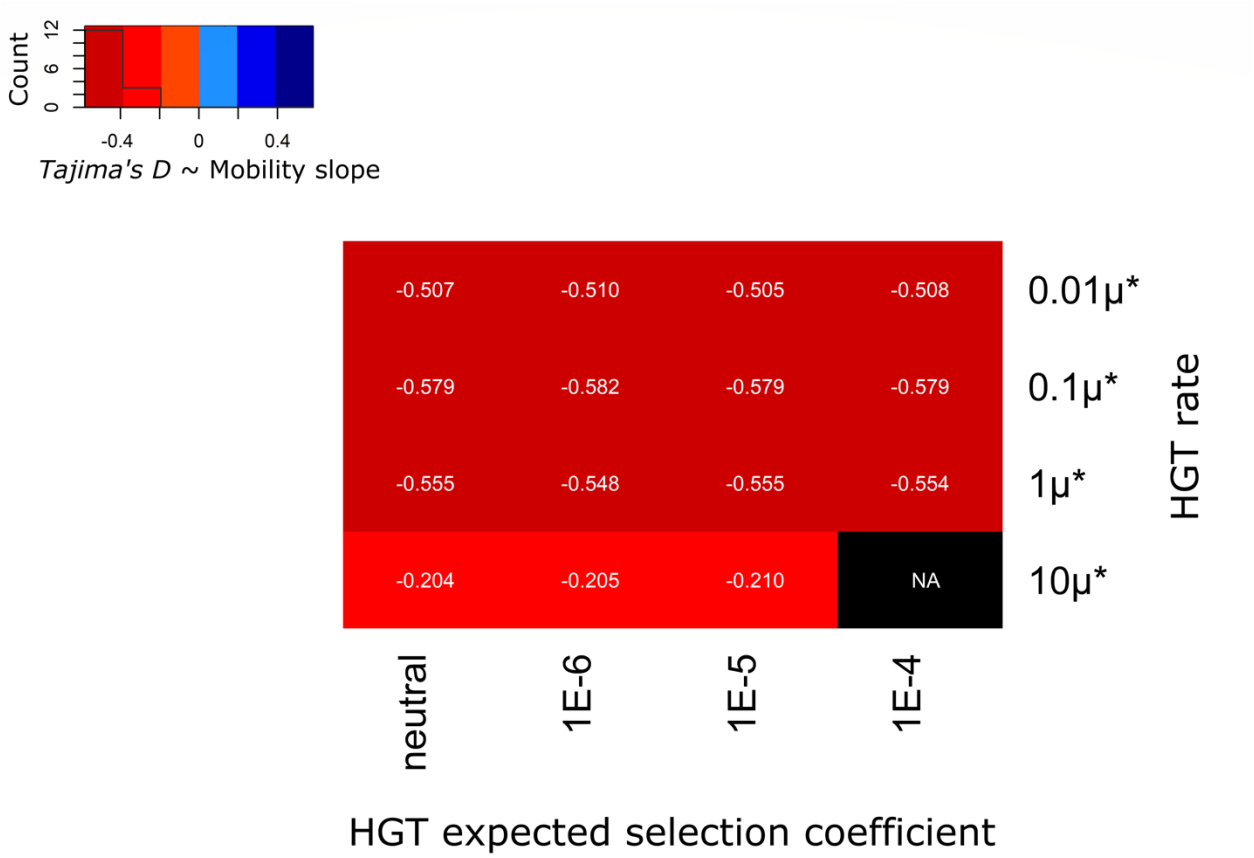


Figure 3.S11 Mobile genes nucleotide diversity is more influenced by HGT rate than HGT selection coefficient in simulations

This figure shows the heatmap of $Tajima's\ D \sim Gene\ Mobility$ average slope across simulations. Regression p-values were obtained through a *t*-test. Data standardization was performed before each regression to respect *t*-test assumption of normality. The slope is indicated in white in the heatmap cells and its sign is represented by the color, i.e. red represents a negative slope, blue represents a positive slope and black represent non-significant results (NA). Across simulations, the slope of the regression varies more because of HGT rate changes than HGT fitness effect. This was confirmed by a K-S test (p-value < 0.05). The test was realized with 160 simulations data and the average slopes were calculated over the 10 replicates of each of the 16 simulation sets (Only the significant slopes were considered). Simulation sets differ by HGT rate or HGT fitness effect: HGT rate $\in [0.01\mu^*, 0.1\mu^*, 1\mu^*, 10\mu^*]$ and HGT $\lambda \in [\text{neutral}, 10^4, 10^5, 10^6]$, where λ is the rate parameter of the exponential distribution of HGT selection coefficient and μ^* is the mutation rate (Methods). The expected selection coefficient corresponds to $\frac{1}{\lambda}$. We increased Prokaryotic mutation

rate μ^* up to the order of 10^{-7} mutations per site per generation for practical reason and made sure it does not change genome size equilibrium (Methods and **Figures 3.S10 p.124**). Each simulation included 5000 cells, 10 species, 500 genes per cells at equilibrium and start, as well as a simulation time of 10^5 generations.

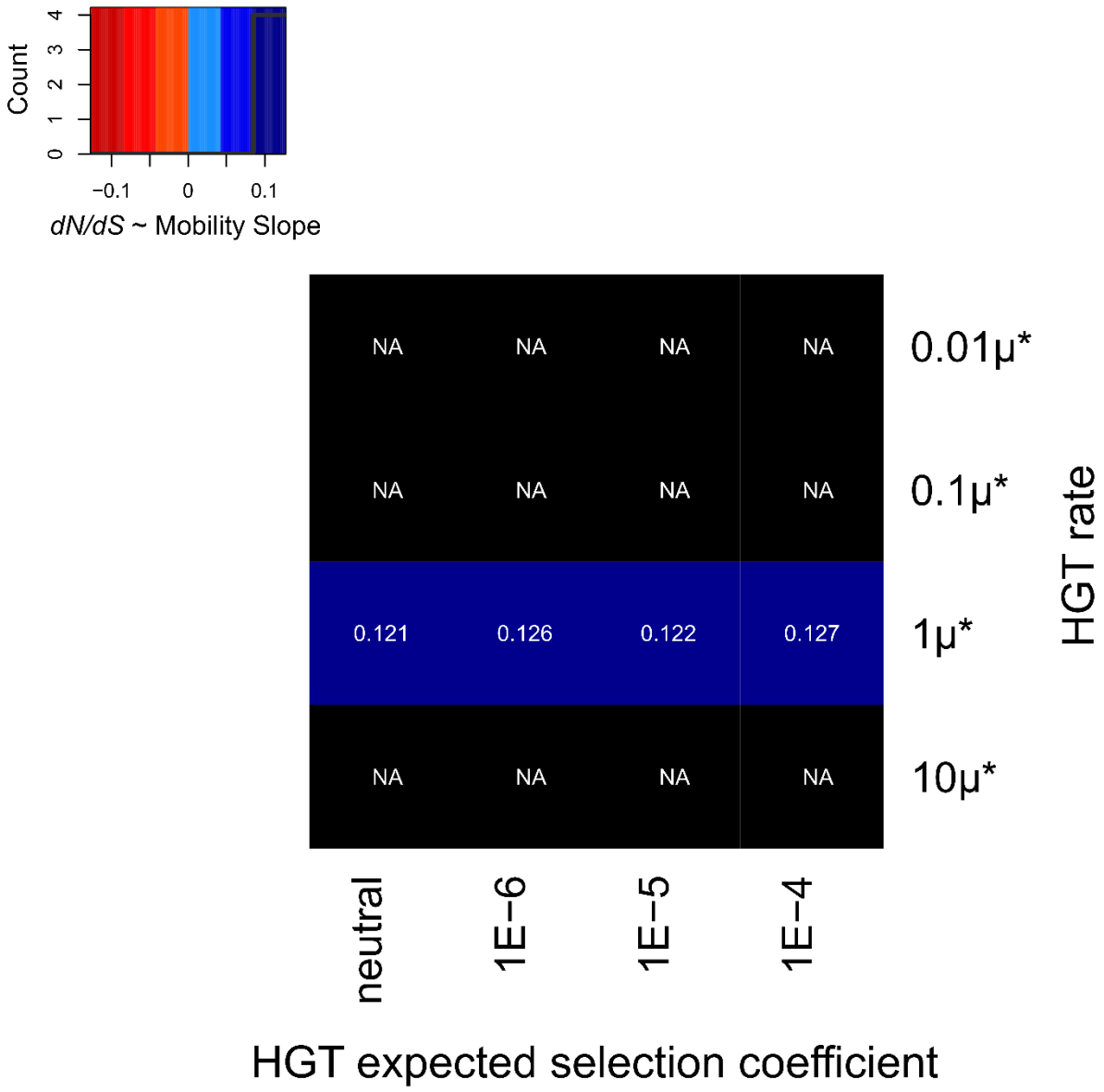


Figure 3.S12 *dN/dS* correlates weakly with gene mobility in simulations

Heatmap of $dN/dS \sim Gene\ Mobility$ average slope across simulations Regression p-values were obtained through a *t*-test. Data standardization was performed before each regression to respect *t*-test assumption of normality. The slope is indicated in white in the heatmap cells and its sign is

represented by the color, i.e. red represents a negative slope, blue represents a positive slope and black represent non-significant results (NA). In the simulation for which the correlation is significant, the results show that *dN/dS* tends to be positively correlated with Gene mobility. However, the correlation is not reproducible across most of the simulations and the average adjusted R^2 was 0.01, which is very weak. The simulations included 5000 cells, 10 species, 500 genes per cells at equilibrium and simulation time of 10^5 generations.

3.4 Conclusion

Pangenome evolution has been studied primarily on long evolutionary time scales by comparing relatively distantly related genomes. Studies of these long time scales have largely concluded, although with some debate, that pangenomes are predominantly adaptive – that selection plays a bigger role in pangenome evolution than drift. Here we have refocused the study of pangenome evolution to shorter time scales, that is within individual gut microbiomes in which gene transfer events likely occurred within a human lifespan. Based on microbiome data from a Fiji cohort, we found that mobile gene sequence evolution is more influenced by selective pressures at the level of gene function than at human host level. Of course, there were many unmeasured human host factors which could impose selective pressures that we were unable to study. However, complementary evolutionary simulation results showed that mobile genes need not provide any special adaptive value to their human host or microbial genomes in order to recapitulate the qualitative patterns of molecular evolution observed in the real data.

These observed patterns of molecular evolution based on population genetic metrics provide clues about the balance of evolutionary forces acting on mobile genes in microbiomes within a human lifespan. We found that most genes accumulate low-frequency mutations as they spread within and between bacterial species. One interpretation of this result is that most mobile genes are under purifying selection to maintain a conserved function, even as they spread across species, such that most mutations are deleterious and kept at low frequency. Another non-exclusive interpretation is that low-frequency mutations could also represent rapid spread of a gene, before mutations are able to rise to higher frequency. In contrast, a minority of genes involved in few specific cellular functions, such as defense mechanisms (COG category V), accumulate intermediate frequency alleles as they spread in new species, possibly due to negative frequency-dependent selection within species and/or fixation of beneficial mutations within some species but not others. Further investigation will be needed to explore the nature of these variable selective pressures.

Similarly to Bobay and Ochman (2018), we observed a very weak correlation between gene mobility and *dN/dS*, which measures selection in protein-coding regions. Bobay and Ochman (2018) attributes this trend to a nearly neutral model of pangenome evolution, *i.e.* drift-barrier evolution. This assumption that most accessory genes are slightly beneficial can explain why a

mixture of neutral and adaptive patterns are evident throughout our analysis. However, further work is needed to test the validity of this model in additional datasets.

Thus, pangenome evolution is the product of a fine balance between drift and selection, which can shift depending on the time scale and level of biological organization, from gene to genome to community. In the gut microbiome of a single person, the time scale of evolution may be too short to easily resolve the balance between drift and selection. Indeed, on very short time scales during which mutations could still be segregating and HGT occurs more frequently than mutation fixation, slightly adaptive genes that have been recently transferred could be largely influenced by drift because of their small N_e, such that their adaptiveness could be effectively detected only on long time scales, while drift might decide their fate on shorter time scales. In this context, it is not surprising that simulations identified HGT rate, but not selective coefficients, as an important driver of mobile genes sequence evolution on short time scales. This model seems to fit some other bacterial genomic datasets (Bobay & Ochman, 2018; Gardon et al., 2020) but awaits formal testing. Finally, we suggest that future work on pangenome evolution should try to understand what factors control shifts in the drift-selection balance and its interplay with species ecology (N_e, species lifestyle, etc.) and gene ecology (*i.e.* gene function, to what extent are genes selfish or cooperative within a genome, etc.), which is probably more informative than simply settling for either an adaptive model or a non-adaptive model.

3.5 Methods

3.5.1 Population genetics of Fijian gut microbiome mobile genes

The Fiji Community Microbiome project provides open access to metagenomes from the gut microbiomes of 176 individuals. For each of these individuals, we mapped metagenomic sequence reads to a set of 37 853 mobile genes previously defined as follows from bacterial whole genome sequences from the Human Microbiome Project (HMP) and FijiCOMP. To be considered mobile, pairs of genes 500bp or longer had to share >99% nucleotide identity between isolate or single-cell genomes with <97% identity in the 16S rRNA gene (Brito et al., 2016). This procedure selects nearly identical genes present in distinct species or genera as candidates for very recent HGT, likely within an individual gut microbiome (Brito et al., 2016; Smillie et al., 2011). An additional filter was applied to remove potential false-positive HGT events from highly conserved ribosomal proteins, and to keep only reads that aligned with 99% identity across >= 50% of their own length (Brito et al., 2016). From the mappings, we used Anvi'o to report Single Nucleotide Variants (SNVs) (--min-coverage-for-variability 10 --min-contig-length 50), followed by a pipeline to compute population genetics metrics (θ_π, θ_w, *dN/dS* and *Tajima's D*) based on the SNVs. The pipeline scripts are available at https://github.com/arnaud00013/Fiji_Mobile_Gene_Specific_PopGen_scripts. The Anvi'o SNV calling module(Eren et al., 2015) has the advantage of being fast and simple to use, can be executed in parallel (High-Performance Computing), and has filters to control minimum gene coverage or mutation frequency. For each sample mapping, a gene was retained if its mean site depth was >= 10. Only one sample was excluded for having less than 500 genes passing the site depth filter, reducing the sample size to 175 metagenomes. Among all samples, 7990 unique genes were conserved after applying the site depth filter. Finally, mobile gene COG annotations, available in the FijiCOMP data (http://fijicomp.bme.cornell.edu//), were used to define two level of gene functions: COG gene family (which is more specific), and COG category (which is more general).

3.5.2 Detecting selection by *dN/dS*

dN/dS is the non-synonymous to synonymous mutations per site ratio. Different methods have been developed to estimate *dN/dS* with the common purpose of inferring selection in protein-coding genes (Spielman & Wilke, 2015). More precisely, *dN/dS* can detect purifying selection (*dN/dS*<1),

neutral evolution (*dN/dS* ≈ 1) and positive selection (*dN/dS* > 1). Because we are working with metagenomic gene variants, we defined our own estimator of *dN/dS*:

$$\frac{\widehat{dN}}{dS} = \frac{Nb_{nsm}/Nb_{nss}}{Nb_{sm}/Nb_{ss}} \quad \text{Éq. 1}$$

where Nb_{nsm} is the number of non-synonymous mutations (SNVs), Nb_{nss} is the number of non-synonymous sites, Nb_{sm} is the number of synonymous mutations (SNVs), and Nb_{ss} is the number of synonymous sites.

3.5.3 Measuring mobile genes nucleotide diversity at metagenomic level

Because mobile genes are by definition present in multiple species, we calculated population genetic metrics based on all reads from a metagenome that map to a particular mobile gene. Based on these mapped reads, we calculated *Tajima's D* (Tajima, 1989), which measures the difference between average per-site pairwise nucleotide differences (θ_π) and the normalized number of polymorphic sites (θ_w) :

$$D_{Tajima} = \frac{\theta_\pi - \theta_w}{\sqrt{\widehat{Var}(\theta_\pi - \theta_w)}} \quad \text{Éq. 2}$$

where the $\widehat{Var}$ denotes the expected sampling variance of $(\theta_\pi - \theta_w)$. For each sample, we estimated mobile gene nucleotide diversity from sequence variants detected in the mapping between metagenomic reads and mobile gene reference sequence from FijiCOMP as follows:

$$\widehat{\theta_\pi} = \frac{Nb\_reads\_pwdiff}{\sum_{i=1}^{n} \langle \frac{c_i}{2} \rangle} \quad \text{Éq. 3}$$

where n is the gene length, c_i is the depth of the site i of the gene and Nb_reads_pwdiff is the number of pairwise nucleotide differences, and

$$\widehat{\theta_w} = \frac{S}{a1} \quad \text{Éq. 4}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$ Éq. 5

where a_1 is a normalizing factor that represents the sample size (n). Usually, *Tajima's D* is estimated from a multiple alignment between gene alleles. The sample size used to estimate the normalizing factor a1 is the number of alleles. Here we use the average depth of coverage at polymorphic sites as an estimator of the sample size n.

3.5.4 Effect of gene mobility on metagenomic coverage

We operationally defined gene mobility as the number of single-cell genomes in which a mobile gene was found and tested if this metric behaves as expected in explaining gene frequencies in metagenomes. More precisely, we correlated gene mobility with metagenomic coverage with the expectation that more mobile genes occur in multiple species and should thus be more deeply covered by metagenomic sequence reads. Linear regression analyses and *t*-tests were calculated using the R function *summary.lm()* (RCoreTeam, 2019). Data standardization was performed before each regression to respect the *t*-test's assumption of normality. The distributions of the significant correlations adjusted R2 obtained from the t-tests converged with the ones from a non-parametric permutational ANOVA (K-S test p-value > 0.05) (Anderson, 2001; Wheeler & M., 2016). The results from the *t*-tests realized on the standardized data are presented instead of the results from the permutational ANOVA realized on the raw data because standardization allows to remove units and thus facilitate the comparison between the Fiji data and the simulation data. Indeed, standardization allows to remove the units and make an easier comparison between the Fiji dataset, in which gene abundance is estimated by coverage, and the simulations, in which gene abundance is estimated by the number of gene copies.

3.5.5 Assessing variation in sequence evolution across genes and across individuals

To determine whether mobile gene evolution is driven more by gene-specific factors or by host attributes, we first compared the variation of mobile genes nucleotide diversity (and other population genetic metric described above) across genes vs. across samples through the Kolmogorov-Smirnov test (KS test). The KS test involves a statistic *D*, which measures the maximal distance between a pair of cumulative distributions. We downsampled the mobile genes

to the same size as the number of samples to avoid the potential bias due to different sized datasets and repeated this for a total of 999 resamples. We performed this series of KS test with the function ks.test() from the R package *stats* (RCoreTeam, 2019).

3.5.6 Gene function and human host (individual) attributes as predictors of mobile genes evolution

To determine whether mobile gene evolution is driven more by gene function or host attributes, we performed linear regressions between a continuous response variable and a qualitative/categorical explanatory variable, which we will refer as a factor. Regressions between a quantitative continuous variable, e.g. *Tajima's D*, and a factor, e.g. gene function family, requires transforming the factor as it cannot be integrated into a regression equation in its original form (RCoreTeam, 2019). We therefore used the R contrast function *constr.sum()* to transform factors (RCoreTeam, 2019). This transformation allows the regression coefficients to represent how each level/state of the factor differ. Then, we assess the significance of the regression model with a non-parametric (permutational) ANOVA (Anderson, 2001). This test makes random permutations of the response variable between the different groups/levels of the factor and estimates the *p*-value as the proportion of permutations with an F-statistic greater than or equal to that observed in the real (unpermuted) data. This test was implemented in the R library *lmPerm* (v.2.1.0) (RCoreTeam, 2019).

For host attribute correlations with population genetic metrics, we focused on 172 samples with available metadata. Metadata about these samples were extracted from Brito et al. (2016) and NCBI accession numbers of the corresponding stool metagenomes are publicly available at http://fijicomp.bme.cornell.edu//data/FijiCOMPmetagenomicsamples.xlsx. Mobile genes selected for this analysis needed to respect the following conditions: (1) the gene should have at least 10X coverage to limit the impact of sequencing errors, and (2) mobile gene should have less than 30% missing values across samples, for a total of 1333 tested genes.

As for linear regressions between population genetics metrics and gene families, we selected genes based on the following set of conditions : (1) the gene should have at least 10X coverage to limit the impact of sequencing errors, (2) the gene should have available gene family annotations, which come from COG, KEGG, TIGRFAM , PFAM or dbCAN databases (Brito et al., 2016) (3) the gene family should be represented by at least 2 genes within the dataset and (4)

the mobile gene should have less than 30% missing values across samples, for a total of 512 tested genes. The first two filters are the basic requirements for doing these regressions analyses. However, the 3rd and 4th filters were chosen respectively to avoid the effects of small sample size for gene families that are underrepresented in the dataset, and to handle missing values caused by gene absence across sample or genes with low coverage in gut metagenomes.

3.5.7 Effect of HGT on sequence evolution

To determine the impact of HGT on mobile gene sequence evolution, multiple linear regressions were performed. In these multiple linear regressions, coverage, Gene Mobility – the number of species in which a mobile gene has been identified when looking for HGT events – and gene length were the explanatory variables and the various population genetic metrics were the response variables. We used the lm() function in R to remove collinearity with QR-decomposition/Gram-Schmidt orthogonalization. Thus, it is possible to assess the effect of Gene Mobility on each population genetics metrics while controlling for the effect of potential confounders like coverage and gene length. The significance of the multiple linear regression was evaluated with the F-test of the R function summary.lm() (RCoreTeam, 2019). For each response variable Y tested (θ_π, θ_w, *dN/dS* and *Tajima's D*), there are two regression models:

$$Y \sim Gene\ Mobility \quad \text{Éq. 6}$$

$$Y* \sim Gene\ Mobility + Coverage + Gene\ length \quad \text{Éq. 7}$$

The asterisk represents the fact that the regression controls for the effects of coverage and gene length, which increase the chance of observing sequencing errors. The adjusted R^2 of a correlation represents the proportion of variable Y variance that is explained by the regression model with a correction for the number of explanatory parameters included in the model (k) and the sample size (n):

$$adjusted\_R^2 = 1 - \frac{\left(SSres/_{n-k-1}\right)}{\left(SStotal/_{n-1}\right)} \quad \text{Éq. 8}$$

where SSres is the residuals sum of squares and SStotal is the fitted data Sum of Squares. The type of correlation (positive or negative) can be determined by the regression coefficient. The

reproducibility of the regressions was measured by the number of samples in which the correlation is significant.

For the simple linear regression, the p-values were obtained using the *t*-test of the R function summary.lm() (RCoreTeam, 2019). Data standardization was performed before each regression to respect the *t*-test's assumption of normality. The distributions of the significant correlations adjusted R^2 obtained from the t-tests converged with the ones from a non-parametric permutational ANOVA (Anderson, 2001; Wheeler & M., 2016). The results from the *t*-tests realized on the standardized data are presented instead of the results from the permutational ANOVA realized on the raw data because standardization allows to remove units and thus facilitate the comparison between the Fiji data and the simulation data. Indeed, in the Fiji dataset, population genetics metrics were estimated from metagenomic mapping and mobility was estimated from single-cell genomes contrarily to the simulation.

3.5.8 Variation across COG categories

To assess how the relationships between gene mobility and *Tajima's D* or coverage varied across COG categories, we considered 22 COG categories (Tatusov et al., 2000). We then used linear mixed models, through the R package lme4, to study the effect of gene mobility on coverage and *Tajima's D* across COG categories (Bates, Mächler, Bolker, & Walker, 2015). A linear mixed model allows to build a linear model between the response variable and the fixed effects while controlling for random effects. In the regression model, fixed effects are explanatory variables for which we want to know the relationship with the response variable. Random effects are grouping factors that explain random variance of the relationship between the response variable and the fixed effects across a finite number of different groups. To control for random effects, the algorithm builds a linear model for each group. In the two regression models, $COG\ category$ and $Sample$ were included as random effects:

$$Coverage \sim Mobility\ +\ COG\ category\ +\ Sample \qquad \text{Éq. 9}$$

$$Tajima's\ D \sim Mobility\ +\ COG\ category\ +\ Sample \qquad \text{Éq. 10}$$

We can then test the significance of $COG\ category$ for the regression model using a permutation ANOVA (Anderson, 2001). The advantage of such test is that it is non-parametric, making no assumptions about the distribution underlying the data. For both regressions, we conducted 99,999 permutations of the response variable between COG categories and then calculated the F-statistic of the regression after each permutation. Next, we calculated the F-statistic of the original regression and calculated the *p*-value as the proportion of permuted data regressions that gave an F-statistic greater than or equal to the F-statistic from the real (non-permuted) data.

Additionally, using the R function *anova()*, we performed likelihood ratio tests between each linear mixed model and their nested models to test the significance of each random factor, i.e. $COG\ category$ and $Sample$ (Crainiceanu & Ruppert, 2004; RCoreTeam, 2019). Each nested model was obtained by removing one random factor at a time, thus creating two nested models per response variable Y:

$$Y \sim Mobility\ +\ COG\ category \quad \text{Éq. 11}$$

$$Y \sim Mobility\ +\ Sample \quad \text{Éq. 12}$$

The likelihood ratio test compares the likelihood of a nested model to the likelihood of the full linear mixed model, with the assumption that the test statistic follows a Chi-square distribution. Thus, we can create each nested model by the removal of a single random factor from the full linear mixed model and assess the significance of both random factors using a *p*-value from the Chi-square distribution (Crainiceanu & Ruppert, 2004).

3.5.9 Simulation of pangenome evolution

We simulated Sela, Wolf and Koonin's prokaryotic genome size evolution model with few changes, using the SodaPop simulation tool (Gauthier et al., 2019; Sela et al., 2016). In this model, the selective advantage of gene gain, *i.e.* the advantage of having $x+1$ genes instead of x genes, depends of the genome size, which is measured by the number of genes in the genome (x). Selection coefficients for gene loss have the opposite sign as gene gain; thus, gene gain is slightly beneficial while gene loss is slightly deleterious (Sela et al., 2016). The selection coefficient of gene gain and gene loss can thus be described by the following formula:

$$s_{gain}(x) = a + b \cdot x = -s_{loss}(x) \quad \text{Éq. 13}$$

where s_{gain} is the selection coefficient of gene gain through HGT, a is a constant input parameter of the simulation that allows to improve the fit of the linear expression with the real data, b is a constant input parameter that represents the benefit or cost associated with the gain of a single gene, x represents genome size (number of genes), and s_{loss} is the selection coefficient of gene loss. We modified this formula to simulate a model where each gene has its own constant selective advantage regardless of genome size (x). To do so, we only needed to set the condition **b=0**. This change allowed us to reproduce the shape of gene mobility distribution in simulation (**Figure 3.S1 p.51**). In this case:

$$s_{gain} = a = s_{gene} = -s_{loss} \quad \text{Éq. 14}$$

where s_{gene} ~ Exp(λ), λ is an input parameter of the simulation, and 1/λ represents the expected value of the exponential distribution of selection coefficients.

In the model, genome size (x) influence gene gain rate and gene loss rate. Indeed, the more genome size increases, the more gene gain rate decreases, and the more gene loss rates increases to find an equilibrium around a certain genome size x_0. Therefore, when genome size (x) is smaller than genome size at equilibrium (x_0), the cell has a higher probability of gene gain than loss. To consider the stochastic component of evolution, the cells and genes that are involved in each gain or loss events are randomly selected. Also, the number of gain or loss events are drawn from a Poisson distribution with the gain and loss rates as follows:

$$G_{rate} \sim Poisson(\lambda = s' \cdot x^{\lambda^+}) \quad \text{Éq. 15}$$

$$L_{rate} \sim Poisson(\lambda = r' \cdot x^{\lambda^-}) \quad \text{Éq. 16}$$

where G_{rate} is the gain rate, i.e. the number of gene gain events per generation, L_{rate} is the loss rate, i.e. the number of gene loss events per generation, and r', s', λ^+ and λ^- are simulation input parameters that allow to tune the gain and loss rates.

We implemented this model in the SodaPop software. which simulates a Wright-Fischer process for asexual populations (Gauthier et al., 2019). In SodaPop, the mutation model is equivalent to Jukes-Cantor in which all single nucleotide occur at the same constant rate (Jukes & Cantor, 1969). We also implemented a distribution of non-synonymous mutation fitness effect in which 30% of mutations are lethal, as previously reported in literature (Eyre-Walker & Keightley, 2007), and 70% are drawn from a normal distribution, $N(\mu=-0.02, \sigma=0.01)$. Synonymous mutations are all considered neutral unless the user provides data on species codon usage and the related fitness effects. SodaPop also offers flexibility in the initial setup of the simulation (Gauthier et al., 2019). We created scripts to facilitate the creation of the simulation starting conditions (https://github.com/arnaud00013/SodaPop/tree/Sodapop-pev/tools). The scripts allow to define each species abundance, gene content, and to define the genes that are mobile (https://github.com/arnaud00013/SodaPop/blob/Sodapop-pev/tools/Setup_SodaPop_with_PEV.py). Mobile genes can be transferred and lost while core genes and accessory genes (defined at the start of the simulation) can only be lost. For each set of simulations sharing the same input parameters, we ran 10 replicates. Each simulation included 5000 cells, 10 species, 500 genes per cells at equilibrium and a simulation time of 10^5 generations and a timestep of 10^4 generations to save simulation data. Population size is small in simulation because of hardware memory limitations. To avoid undesirable effects, like Muller's Ratchet, we maintained species abundance constant. We also established a relatively high mutation rate on the order of 10^{-7} mutations per site per generation to compensate for small population sizes. Genome size equilibrium was reached for every simulation and the model is thus robust to the initial conditions (**Figures 3.S10 p.124**). The software is available on GitHub (https://github.com/arnaud00013/SodaPop).

Acknowledgements

We would like to thank Compute Canada for allocated resources, and Louis-Marie Bobay and Gavin Douglas for constructive comments. BJS was supported by an NSERC Discovery Grant. AWRS acknowledges funding from a Canada Research Chair Tier 2 and an NSERC Discovery Grant. AN is supported by an FRQNT scholarship.

3.6 References

1 Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435-439, doi:10.1038/nature18927 (2016).

2 Valdes, A. M., Walter, J., Segal, E. & Spector, T. D. Role of the gut microbiota in nutrition and health. *BMJ* **361**, k2179, doi:10.1136/bmj.k2179 (2018).

3 Garud, N. R. & Pollard, K. S. Population Genetics in the Human Microbiome. *Trends Genet.* **36**, 53-67, doi:10.1016/j.tig.2019.10.010 (2020).

4 Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.* **23**, 598-605, doi:10.1016/j.tim.2015.07.006 (2015).

5 Jiang, X., Hall, A. B., Xavier, R. J. & Alm, E. J. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLoS One* **14**, e0223680, doi:10.1371/journal.pone.0223680 (2019).

6 McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040, doi:10.1038/nmicrobiol.2017.40 (2017).

7 Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11399-11407, doi:10.1073/pnas.1614083113 (2016).

8 Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J* **8**, 1553-1565, doi:10.1038/ismej.2014.60 (2014).

9 Moulana, A., Anderson, R. E., Fortunato, C. S. & Huber, J. A. Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents. *mSystems* **5**, doi:10.1128/mSystems.00673-19 (2020).

10 Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J* **11**, 1719-1721, doi:10.1038/ismej.2017.36 (2017).

11 Bobay, L. M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153, doi:10.1186/s12862-018-1272-4 (2018).

12 Takeuchi, N., Cordero, O. X., Koonin, E. V. & Kaneko, K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* **13**, 20, doi:10.1186/s12915-015-0131-7 (2015).

13 Shapiro, B. J. The population genetics of pangenomes. *Nat Microbiol* **2**, 1574, doi:10.1038/s41564-017-0066-6 (2017).

14 Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E. & Koonin, E. V. Two fundamentally different classes of microbial genes. *Nat Microbiol* **2**, 16208, doi:10.1038/nmicrobiol.2016.208 (2016).

15 Hehemann, J. H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908-912, doi:10.1038/nature08937 (2010).

16 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

17 Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241-244, doi:10.1038/nature10571 (2011).

18 Vogan, A. A. & Higgs, P. G. The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biol. Direct* **6**, 1, doi:10.1186/1745-6150-6-1 (2011).

19 Corander, J. *et al.* Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol* **1**, 1950-1960, doi:10.1038/s41559-017-0337-x (2017).
20 Domingo-Sananes, M. R. & McInerney, J. O. Selection-based model of prokaryote pangenomes. *bioRxiv*, 782573, doi:10.1101/782573 (2019).
21 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).
22 Koonin, E. V. & Wolf, Y. I. Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* **11**, 487-498, doi:10.1038/nrg2810 (2010).
23 Gardon, H., Biderre-Petit, C., Jouan-Dufournel, I. & Bronner, G. A drift-barrier model drives the genomic landscape of a structured bacterial population. *Mol. Ecol.*, doi:10.1111/mec.15628 (2020).
24 Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263-273, doi:10.1038/nrmicro3218 (2014).
25 Gauthier, L., Di Franco, R. & Serohijos, A. W. R. SodaPop: a forward simulation suite for the evolutionary dynamics of asexual populations on protein fitness landscapes. *Bioinformatics* **35**, 4053-4062, doi:10.1093/bioinformatics/btz175 (2019).
26 Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610-618, doi:10.1038/nrg2146 (2007).
27 Bachtrog, D. & Gordo, I. Adaptive evolution of asexual populations under Muller's ratchet. *Evolution* **58**, 1403-1413, doi:10.1111/j.0014-3820.2004.tb01722.x (2004).
28 Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319, doi:10.7717/peerj.1319 (2015).
29 Spielman, S. J. & Wilke, C. O. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol* **32**, 1097-1108, doi:10.1093/molbev/msv003 (2015).
30 A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2019).
31 Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32-46, doi:10.1111/j.1442-9993.2001.01070.pp.x (2001).
32 Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33-36, doi:10.1093/nar/28.1.33 (2000).
33 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1-48 (2015).
34 Crainiceanu, C. & Ruppert, D. Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **66**, 165-185 (2004).
35 Jukes, T. H. & Cantor, C. R. in *Mammalian protein metabolism* Vol. III (ed Elsevier) Ch. 24, 21-132 (Academic Press, 1969).

4 Outil de simulation évolutive

Ce chapitre a pour objectif de présenter en détail le fonctionnement du module d'évolution pangénomique que nous avons implémenté dans SodaPop (Gauthier et al., 2019).

4.1 Simulation de l'évolution pangénomique

Le module d'évolution pangénomique de SodaPop simule explicitement le modèle d'évolution de procaryotes de Sela, Wolf et Koonin (Gauthier et al., 2019; Sela et al., 2016) avec certains changements. Dans ce modèle, l'avantage sélectif du gain de gène, c'est-à-dire l'avantage d'avoir *x+1* gènes au lieu de x gènes, dépend de la taille du génome, qui est mesurée par le nombre de gènes dans le génome (*x*). Le coefficient de sélection de la perte de gène a un signe opposé à celui du gain de gène et plus précisément, le gain de gène est légèrement bénéfique tandis que la perte de gène est légèrement délétère (Sela et al., 2016). Le coefficient de sélection du gain de gène et de la perte de gène peuvent ainsi être décrits par la formule suivante:

$$s_{gain}(x) = a + b \cdot x = -s_{loss}(x) \qquad \text{Éq. 13}$$

où s_{gain} est le coefficient de sélection du gain de gène par THG, a est un paramètre d'entrée constant de la simulation permettant de contrôler la relation linéaire entre s_{gain} et x, b est un paramètre d'entrée constant qui représente l'avantage ou le coût associé au gain d'un seul gène, x représente la taille du génome (le nombre de gènes dans le génome d'une espèce) et s_{loss} est le coefficient de sélection de la perte de gène. Nous avons modifié cette formule pour simuler un modèle où chaque gène a son propre avantage sélectif constant, exponentiellement distribué et indépendant de la taille du génome (*x*), c'est-à-dire que **b=0**. Ce changement permet aux simulations de reproduire la forme de la distribution de la mobilité des gènes dans notre jeu de données réelles (**Figure 3.S1 p.51**); notez que l'espérance d'une distribution exponentielle est 1/λ. Ainsi:

$$s_{gain} = a = s_{gene} = -s_{loss} \qquad \text{Éq. 14}$$

où $s_{gene} \sim Exp(\lambda)$, λ est un paramètre d'entrée de la simulation et $1/\lambda$ représente l'espérance de la distribution du coefficient de sélection du THG.

De plus, dans le modèle, la taille du génome (x) influence les taux de gain et de perte de gène. En effet, à mesure que la taille du génome augmente, le taux de gain de gène diminue et le taux de perte de gène augmente pour maintenir un équilibre autour d'une certaine taille de génome x_0 (Sela et al., 2016). La taille du génome à l'équilibre x_0 représente donc la taille du génome à laquelle le taux de gain est égal au taux de perte. Par conséquent, lorsque la taille du génome (x) est inférieure à x_0 la cellule devrait avoir une probabilité plus élevée de gagner des gènes que de perdre des gènes. Quant au cas où la taille du génome (x) est supérieure à x_0, la cellule devrait avoir une probabilité plus élevée de perdre des gènes que d'en acquérir de nouveaux. Pour tenir compte de la composante stochastique de l'évolution, les cellules et les gènes impliqués dans chaque événement de gain ou de perte sont sélectionnés au hasard. De plus, le nombre d'événements de gain ou de perte est tiré d'une distribution de Poisson avec le taux de gain ou de perte représentant l'argument de la distribution:

$$G_{rate} \sim Poisson(\lambda = s' \cdot x^{\lambda^+}) \quad \text{Éq. 15}$$

$$L_{rate} \sim Poisson(\lambda = r' \cdot x^{\lambda^-}) \quad \text{Éq. 16}$$

où G_{rate} est le taux de gain de gène, c'est-à-dire le nombre d'événements de gain de gène par cellule par génération, L_{rate} est le taux de perte de gène, c'est-à-dire le nombre d'événements de perte de gène par cellule par génération, x représente la taille du génome en nombre de gènes et *r', s',* λ^+ et λ^- sont des paramètres d'entrée de la simulation qui permettent de contrôler G_{rate} et L_{rate}. Il est également important de noter que le THG n'est pas limité aux cellules de 2 espèces différentes, mais pourrait également se produire, plus rarement, au sein d'une même espèce comme c'est le cas entre des souches d'une même espèce bactérienne (Adato, Ninyo, Gophna, & Snir, 2015).

De plus, nous avons choisi d'implémenter ce modèle dans le logiciel SodaPop, car il permet de simuler des mutations et un processus de Wright-Fischer pour les populations asexuées (Gauthier et al., 2019). Dans SodaPop, le modèle de mutation actuel est équivalent au modèle Jukes-Cantor dans lequel tous les changements nucléotidiques se produisent à la même vitesse constante (Jukes & Cantor, 1969). Nous avons également implémenté une distribution des effets des mutations non synonymes sur le succès reproducteur des cellules où 30% des mutations non synonymes sont létales, tel que reporté précédemment dans la littérature (Eyre-Walker &

Keightley, 2007), et 70% sont tirées de la distribution normale **N(μ = -0,02, σ = 0,01)**. Les mutations synonymes sont considérées comme neutres à moins que l'utilisateur ne fournisse des données sur le biais d'utilisation de codons synonymes (https://github.com/arnaud00013/SodaPop/blob/Sodapop-pev/tools/add_Codon_Usage_data_into_Sodapop_cell_file_from_Kasuza_db_like_file.sh) et son effet sur le succès reproducteur des cellules. SodaPop offre également de la flexibilité, car l'utilisateur peut créer lui-même la configuration initiale de la simulation (Gauthier et al., 2019). Nous avons créé des scripts pour faciliter la configuration initiale de la simulation (https://github.com/arnaud00013/SodaPop/tree/Sodapop-pev/tools). Ces scripts permettent de définir l'abondance de chaque espèce, leur contenu en gène et même si un gène est mobile ou non (https://github.com/arnaud00013/SodaPop/blob/Sodapop-pev/tools/Setup_SodaPop__with_PEV.py). Au début de chaque simulation, le groupe de gènes de la communauté microbienne contient des gènes domestiques qui sont présents dans toutes les espèces et des gènes accessoires qui sont répartis au hasard à travers les espèces. Dans cet ensemble de gènes accessoires, certains sont étiquetés aléatoirement comme étant mobiles et la quantité de gènes mobiles dans le groupe de gènes accessoires est définie par l'utilisateur. Les gènes mobiles peuvent être transférés et perdus tandis que d'autres gènes peuvent seulement être perdus. Pour chaque ensemble de simulations partageant les mêmes paramètres d'entrée, nous avons exécuté 10 répliques pour montrer que les résultats étaient reproductibles. Chaque simulation comprenait 5000 cellules, 10 espèces, 500 gènes par cellule à l'équilibre et un temps de simulation de 10^5 générations avec une sauvegarde des données de simulation à chaque 10^4 générations. Les tailles de population en simulation sont plus petites que ce qui peut être observé en réalité en raison des limites de la mémoire computationnelle (Bobay & Ochman, 2018; Sela et al., 2016). Pour nous assurer que cette limitation n'entraîne pas d'effets indésirables, comme l'accumulation de mutations délétères conduisant à l'extinction d'une espèce, également connue sous le nom de cliquet de Muller (Bachtrog & Gordo, 2004), nous avons maintenu l'abondance des espèces constante. Pour éviter un manque de diversité génétique dans la population simulée en raison de la petite taille de la population, nous avons également augmenté le taux de mutation procaryote jusqu'à l'ordre de 10^{-7} mutations par site par génération. Bien que ces paramètres ne soient pas typiques des procaryotes, les résultats de la simulation sont tout de même pertinents, car ils reproduisent les tendances observées dans le jeu de données réel (**Figures 3.4 p.71**). De plus, l'équilibre de la taille du génome

a été atteint pour chaque simulation (**Figures 3.S10 p.124**). Ainsi, nos résultats de simulation ne dépendent pas des conditions initiales. Le logiciel est disponible sur GitHub (https://github.com/arnaud00013/SodaPop).

4.2 Étapes à suivre pour les simulations d'évolution pangénomique

La version de SodaPop qui inclut le module d'évolution pangénomique, c'est-à-dire Sodapop-pev, peut être téléchargée sur https://github.com/arnaud00013/SodaPop (voir https://louisgt.github.io/SodaPop/_2017/09/05/Setup-and-installation.html pour les consignes d'installation).

1) Créez la configuration initiale de la simulation avec le script https://github.com/arnaud00013/SodaPop/blob/Sodapop-pev/tools/Setup_SodaPop_with_PEV.py. Celui-ci se trouvera dans le répertoire **tools/** du dossier téléchargé **Sodapop-pev/**. Ce script crée les fichiers d'entrée nécessaires à l'exécution des simulations de SodaPop. Tous les fichiers d'en-tête doivent se trouver dans le répertoire **files/headers/** de l'espace de travail principal de SodaPop (voir https://github.com/arnaud00013/SodaPop/tree/Sodapop-pev/files/header). Le script est interactif et demande à l'utilisateur la taille du génome de chaque espèce.

   -Arguments:

   (i) le chemin absolu de l'espace de travail SodaPop

   (ii) le nom du fichier d'en-tête **.gene**

   (iii) le nom du fichier d'en-tête **.cell**

   (iv) le nom du fichier d'en-tête des données de population

   (v) le nombre d'espèces simulées

   (vi) la taille de l'ensemble du groupe de gènes accessoires de la communauté microbienne simulée

   (vii) le nombre de gènes domestiques dans chacun des génomes d'espèce au début de la simulation

   (viii) le nombre de gènes accessoires mobiles

   -Dépendances: Python3 (possible de spécifier la version dans l'en-tête du fichier; python3.6 par défaut)

-Exemple de commande utilisée pour notre article sur l'évolution du pangénome sur de courtes échelles de temps:

python3.6 Setup_SodaPop_with_PEV.py ~/Sodapop-pev/ header.gene header.cell header_pop.dat 10 5000 100 200

2) (OPTIONNEL) Ajoutez les données d'utilisation des codons d'une espèce dans son fichier **.cell** avec le script suivant :
https://github.com/arnaud00013/SodaPop/blob/Sodapop-pev/tools/add_Codon_Usage_data_into_Sodapop_cell_file_from_Kasuza_db_like_file.sh

Ce script prend en entrée le chemin absolu du fichier contenant les données d'utilisation de codon de l'espèce dans le format de la base de données Kazusa (1er argument d'entrée) et ajoute les données dans le fichier **.cell** de l'espèce (2e argument d'entrée). Vous devez exécuter le script pour chaque espèce simulée.

-Arguments:

(i) le chemin absolu du fichier contenant les données d'utilisation de codon de l'espèce dans le format de la base de données Kazusa (voir https://www.kazusa.or.jp/codon/) (Nakamura, Gojobori, & Ikemura, 2000)

(ii) le nom du fichier **.cell** de l'espèce

-Exemple:

./add_Codon_Usage_data_into_Sodapop_cell_file_from_Kasuza_db_like_file.sh ~/Sodapop-pev/E_coli_codon_usage.txt ~/Sodapop-pev/0.cell

3) Exécutez **sodasumm** pour créer le fichier **.snap** initial de la population simulée (Gauthier et al., 2019)(voir https://louisgt.github.io/SodaPop/2017/09/05/Running-a-basic-simulation.html)

-Exemple:

sodasumm ~/Sodapop-pev/files/start/population.dat 0

4) Exécutez la simulation en activant le module d'évolution pangénomique

-Arguments:

--sim-type s permet de définir que le type de simulation est celui où l'utilisateur peut définir une distribution de coefficient de sélection pour les mutations non synonymes

--f 9 permet de sélectionner la fonction de succès reproducteur 9 soit celle qui doit être utilisée pour les simulations d'évolution pangénomique puisqu'elle tient compte des effets sélectifs des mutations, du THG, de la perte de gènes.

--normal --alpha -0.02 --beta 0.01 permet de définir la distribution des coefficients de sélection des mutations non synonymes en tant qu'une distribution normale **N(μ=-0.02, σ=0.01)**

-p ~/SodaPop_pev/files/start/population.snap définit le chemin absolu du fichier **.snap** de l'image initiale de la communauté ou population bactérienne simulée

-g~/SodaPop_pev/files/genes/gene_list.dat définit le chemin absolu vers le fichier de liste de gènes

-o simulation_test1 définit le nom de l'espace de travail de sortie (créé s'il n'existe pas déjà et écrasé s'il existe déjà)

-t 20000 définit le pas de temps, c'est-à-dire le nombre de générations entre 2 fichiers d'image **.snap** de la simulation

-n 5000 définit le nombre de cellules simulées

-m 100001 définit le temps de simulation, c'est-à-dire le nombre de générations simulées incluant la génération 0.

-s 2 permet de sauvegarder les fichiers d'image de simulation **.snap** au format long avec séquence d'ADN

-V active le module d'évolution pangénomique et force l'utilisation de la fonction de succès reproducteur 9, qui tient compte des effets sélectifs des mutations, du THG et de la perte de gène

--exp_rate_s_hgt 1E4 définit le paramètre de taux λ de la distribution exponentielle du coefficient de sélection du THG (où 1/λ est l'espérance de la distribution)

--bForSx 0 définit que **b=0** dans la formule du coefficient de sélection du gain de gène (**Équation 13 p.99**)

--rPrime 7.2E-15 définit r' dans la formule du taux de perte de gène (**Équation 16 p.100**)

--sPrime 56250 définit s' dans la formule du taux de gain de gène (**Équation 15 p.100**)

--lambdaPlus -2 définit λ^+ dans la formule du taux de gain de gène (**Équation 15 p.100**)

--lambdaMinus 5 définit λ^- dans la formule du taux de perte de gène (**Équation 16 p.100**)

-e permet de sauvegarder un fichier de suivi de toutes les mutations apparaissant lors de la simulation

-T permet de sauvegarder un fichier de suivi de tous les événements d'évolution pangénomique (gain et perte de gènes), ainsi qu'un fichier de suivi des paramètres liés à cela (taille du génome, taux de perte et taux de THG)

--execVA permet d'effectuer une analyse de l'évolution des gènes mobiles à la fin de la simulation (**execVA** signifie « exécuter l'analyse des variantes de séquence des gènes mobiles »)

-u 6 définit le nombre de processeurs à utiliser pour l'analyse des variantes de séquence

--simulCUB permet de simuler le biais d'utilisation des codons synonymes

--stdCubDfe définit l'écart type (σ) de la distribution des coefficients de sélection des mutations synonymes, où le coefficient de sélection d'une mutation synonyme suit la distribution $\mathbf{N(\mu = 0, \sigma)}$. Cette distribution permet de définir les mutations synonymes comme neutres en moyenne tout en autorisant les mutations synonymes non neutres. Les coefficients de sélection des mutations synonymes sont tirés de la partie positive de la distribution ($\boldsymbol{s}$**>=0**) s'ils augmentent l'index d'adaptation de codon (CAI) (Sharp & Li, 1987) et de la partie négative de la distribution ($\boldsymbol{s}$**<=0**) sinon.

-Example:

sodapop --sim-type s -f 9 --normal --alpha -0.02 --beta 0.01 -p ~/SodaPop_pev/files/start/population.snap -g ~/SodaPop_pev/files/genes/gene_list.dat -o simulation_test1 -t 20000 -n 5000 -m 100001 -s 2 -V --exp_rate_s_hgt 1E4 --bForSx 0 --rPrime 7.2E-15 --sPrime 56250 --lambdaPlus -2 --lambdaMinus 5 -e -T --execVA -u 6

Pour exécuter d'autres types de simulation, veuillez consulter https://louisgt.github.io/SodaPop/2017/09/05/Command-line-flags.html.

5 Discussion

5.1 Discussion générale

En mesurant des paramètres évolutifs à l'échelle de génomes entiers et en considérant de longues échelles de temps évolutifs, un modèle adaptatif d'évolution semble concorder avec les données génomiques de plusieurs espèces procaryotes (McInerney et al., 2017; Sela et al., 2016). En revanche, il est possible que ce ne soit pas le cas durant des échelles de temps plus courtes. En tenant compte des variations observées entre différents gènes mobiles (Wolf et al., 2016) et en se concentrant sur l'évolution des séquences de gènes mobiles sur de courtes échelles de temps, c'est-à-dire dans le microbiote intestinal d'un individu, ce projet cherche à déterminer quel modèle d'évolution explique le mieux l'évolution pangénomique à court terme.

Tout d'abord, nos résultats soutiennent le fait que nous avons bien adapté les paramètres de génétique des populations à l'étude évolutive des gènes mobiles à l'échelle métagénomique, c'est-à-dire en considérant que ces gènes ne sont pas confinés à une seule espèce. En effet, les paramètres que nous utilisons pour trouver des signatures de forces évolutives et mesurer la diversité génétique des gènes mobiles dans le microbiote intestinal, c'est-à-dire θ_π, θ_w, le *D de Tajima* et *dN/dS*, ont des intervalles de valeur qui convergent avec la littérature (**Figure 3.S4 p.58**). De plus, en estimant la mobilité d'un gène mobile avec le nombre de génomes assemblés dans lequel le gène est impliqué dans un événement de THG récent, nous obtenons la corrélation positive attendue entre l'abondance relative et la mobilité du gène (**Figure 3.1 p.52, Table 3.S1 p.51 et Table 3.S3B p.76**). Les génomes assemblés proviennent souvent de genres taxonomiques différents (Brito et al., 2016), ce qui soutient le fait que la présence d'un gène à travers ces différents génomes est causée par des événements de THG. Malgré le fait que nous observons que cette corrélation est positive, la variation des pressions sélectives entre les gènes mobiles et la sélection négative pourraient être responsable de la réduction de la force de cette corrélation, mais pas suffisamment pour l'aplatir complètement (**Figure 3.1 p.52**).

La corrélation positive observée entre l'abondance relative du gène et sa mobilité ne permet pas de favoriser un modèle d'évolution pangénomique particulier puisque cela est attendu tant par un modèle adaptatif que par un modèle non adaptatif. Nous nous sommes donc appuyés sur d'autres

observations pour caractériser l'équilibre entre la sélection et la dérive génétique. Puisque le THG semble être en moyenne légèrement bénéfique (Bobay & Ochman, 2018; Sela et al., 2016) et que certaines familles de gènes mobiles comme les gènes de résistance aux antibiotiques sont sélectionnés dans le microbiote intestinal humain durant le temps de vie d'un individu (Jiang et al., 2019), mon hypothèse de départ était que nous devrions être capables d'observer qu'un modèle adaptatif explique mieux l'évolution pangénomique sur de courtes échelles de temps. Cependant, nos résultats ne valident pas cette hypothèse ou du moins, ils permettent d'apporter une réponse plus nuancée. En effet, les simulations nous ont montré qu'un modèle évolutif simple pouvait récapituler nos observations sur l'évolution pangénomique à court terme sans avoir besoin de considérer que les gènes mobiles ont un effet sélectif sur leur hôte microbien ou humain. Plus précisément, des simulations durant lesquelles le THG n'a pas d'effet sélectif permettent d'expliquer les corrélations observées dans le vrai jeu de données métagénomiques entre la mobilité de ces gènes et les paramètres de génétique des populations, soit θ_π, θ_w et *le D de Tajima* (**Figure 3.4 p.71**). Nous avons aussi observé qu'il y a une sélection pour des fonctions précises dans le microbiote intestinal, comme les fonctions liées aux mécanismes de défense ou au métabolisme secondaire (**Figure 3.5 p.74**), suggérant que ces contraintes sélectives ne s'appliquent pas nécessairement à la majorité des gènes mobiles. De plus, nous avons observé que *dN/dS*, qui mesure la sélection, ne corrèle pas significativement avec le niveau de mobilité des gènes (**Figures 3.S8 p.69 et 3.S12 p.83**). Même si certains de ces résultats peuvent être expliqués de manière plus parcimonieuse par un modèle d'évolution neutre, un modèle d'évolution quasi neutre où les gènes mobiles sont légèrement bénéfiques en moyenne peut mieux expliquer le mélange de patrons d'évolution neutre et adaptative observés tout au long de notre analyse.

Ce projet cherche aussi à évaluer l'impact des attributs de l'hôte du microbiote intestinal sur l'évolution pangénomique durant de courtes échelles de temps évolutif. En comparant des populations humaines ayant divergé il y a des milliers d'années, Brito et ses collaborateurs (2016) ont découvert que la composition en espèces et le groupe de fonctions des gènes mobiles du microbiote intestinal étaient significativement influencés par le mode de vie des individus. Mon hypothèse de départ était que sur de courtes échelles de temps, à l'intérieur d'une seule population humaine et en s'intéressant à l'évolution des séquences de gènes mobiles comme marqueur d'évolution pangénomique, les attributs de l'hôte risquent d'avoir une influence plus faible que la fonction du gène sur l'évolution des gènes mobiles. Nos observations valident cette hypothèse. En

effet, les attributs d'hôte étudié n'expliquent pas significativement l'évolution des gènes mobiles à court terme (**Figures 3.3 p.61 et 3.S5 p.63**). Cela peut être expliqué par le fait que les variations observées entre différents individus d'une même population sont relativement faibles sur de courtes échelles de temps (**Figure 3.2 p.59**). De plus, les mutations que les gènes mobiles acquièrent mettent du temps à se fixer, surtout si ces gènes ont une faible N_e, de sorte que l'effet sélectif des attributs d'hôte est difficilement perceptible à court terme contrairement à ce qui s'observe sur de longues échelles de temps (Brito et al., 2016). Cela concorde avec le fait que les simulations montrent que l'évolution des gènes mobiles est significativement influencée par le taux de transfert horizontal du gène ou autrement dit son niveau de mobilité (**Figure 3.S11 p.82**) et supporte ainsi un modèle d'évolution quasi neutre. Quant à la fonction du gène mobile, cela semble avoir un impact significatif sur l'évolution pangénomique à court terme. En effet, les variations d'évolution de gènes mobiles mesurées à partir des paramètres de génétique des populations sont mieux expliquées par la famille/fonction d'un gène que par les attributs d'hôte (**Figures 3.2 p.59 et 3.3 p.61**).

5.2 Jeu de données

Afin d'étudier l'évolution pangénomique des gènes mobiles sur de courtes échelles de temps, le jeu de données sélectionné semble être approprié puisque l'approche de détection des gènes mobiles utilisée permet d'inférer des événements récents de THG (Brito et al., 2019; Smillie et al., 2011). De plus, la disponibilité d'annotations fonctionnelles sur ces gènes a permis de déterminer que la fonction des gènes est un facteur plus important que les attributs de l'hôte du microbiote. Ces annotations proviennent des bases de données COG, KEGG, TIGRFAM, PFAM ou dbCAN, ce qui permet de renforcer la quantité et la qualité d'information sur ces gènes. Quant à la disponibilité de métadonnées sur les personnes impliquées dans l'étude, cela permet d'évaluer l'impact des attributs d'hôte sur l'évolution pangénomique durant de courtes périodes et observer que leur impact est très faible contrairement à ce qui a été observé sur de plus longues échelles de temps à travers différentes populations humaines (Brito et al., 2016; Yatsunenko et al., 2012; Zhernakova et al., 2016).

Cependant, il est important de noter que dans le jeu de données, les événements de THG ont été détectés à partir de données provenant de seulement 180 génomes assemblés. Par contre, ces génomes représentent probablement les espèces les plus abondantes du microbiote intestinal,

ce qui les rend plus facilement détectables dans les échantillons. Ces espèces abondantes représentent probablement la majorité de la diversité génétique des échantillons, ce qui rend le jeu de données acceptable (Brito et al., 2016). Par la suite, les annotations manquantes, la faible couverture de séquençage de certains gènes mobiles et la rareté ou même l'absence de certains gènes dans la cohorte fidjienne étudiée limitent notre analyse à une fraction du jeu de données. Par exemple, seulement 7990 sur 37 853 gènes mobiles ont une couverture de séquençage moyenne >= 10 dans au moins un des 176 échantillons, une exigence minimale pour les analyses de génétique des populations. De plus, pour les analyses fonctionnelles, nous avons été réduits à 1333 gènes mobiles pour lesquels la famille fonctionnelle était connue et avait plus qu'un représentant dans le jeu de données. Cela peut limiter la portée de nos résultats et c'est pour cela qu'il fût nécessaire d'évaluer la robustesse de nos résultats. Par exemple, même si nous observons que la fonction des gènes a plus d'impact que les attributs d'hôte sur l'évolution pangénomique durant de courtes échelles de temps (**Figures 3.3 p.61 et 3.S5 p.63**), cette corrélation dépend des gènes sélectionnés (**Figure 3.S7 p.67**). Cependant, nous observons qu'il y a une sélection pour des fonctions précises dans le microbiote intestinal, comme les fonctions liées aux mécanismes de défense ou au métabolisme secondaire, suggérant que ces contraintes sélectives ne s'appliquent pas nécessairement à la majorité des gènes mobiles (**Figure 3.5 p.74**). Enfin, il est clair que nous n'étudions pas l'ensemble des attributs d'hôte d'intérêt biologique ou évolutif. Par contre, les facteurs d'hôte étudiés sont ceux qui démontrent un impact significatif sur le contenu en gène du microbiote sur des échelles de temps plus longues, c'est-à-dire en comparant différentes populations humaines (Falony et al., 2016; Yatsunenko et al., 2012; Zhernakova et al., 2016). De plus, les simulations ont montré que l'effet du THG sur l'évolution des séquences de gènes mobiles pouvait être expliquée par un modèle évolutif simple qui n'inclut pas l'avantage sélectif que les gènes mobiles du microbiote pourraient conférer à un hôte ayant des attributs spécifiques. Cela supporte aussi le fait que les facteurs d'hôte ont en général peu d'impact sur l'évolution pangénomique à court terme.

5.3 Perspective

Malgré la quantité de résultats d'intérêt obtenus avec le jeu de données de Brito et al. (2016), des analyses similaires appliquées à d'autres jeux de données métagénomiques, comme celui contenant des données sur des microbiomes intestinaux nord-américains du projet de microbiome humain,

doivent être réalisées. Ces analyses peuvent nous aider à valider nos résultats ou découvrir l'existence de variations à travers diverses populations humaines. De plus, en explorant des jeux de données qui ne sont pas liés au microbiote intestinal humain, il est possible que l'équilibre des forces évolutives varie d'un milieu à l'autre. Par exemple, l'évolution des endosymbiotes ou des agents pathogènes intracellulaires, qui ont une petite taille efficace de population, est généralement modulée par la dérive génétique et cela leur confère de petits pangénomes (Giovannoni et al., 2014). En revanche, la sélection semble jouer un rôle plus important dans les microbes sans hôte qui vivent dans des conditions extrêmes, comme les bactéries hydrothermales (Moulana et al., 2020).

Dans ce contexte, les nouvelles analyses doivent donc aussi s'intéresser à l'interaction entre l'écologie et l'évolution pangénomique. Par exemple, il serait possible de déterminer comment la N_e des gènes mobiles est affecté par l'écologie des espèces et la structure des populations dans lesquels ces gènes se retrouvent. La littérature révèle que l'écologie d'une espèce affecte sa taille efficace de population (McInerney et al., 2017), mais l'influence sur la N_e des gènes mobiles n'est pas clairement détaillée.

De plus, malgré le fait que certains de nos résultats peuvent être expliquée de manière parcimonieuse par un modèle où le THG est neutre, nous avons aussi observé des signatures d'évolution adaptative. Donc, un modèle d'évolution presque neutre (quasi neutre) peut mieux réconcilier nos résultats. Les futurs travaux devraient donc tester la validité du modèle d'évolution presque neutre de barrière-dérive sur de courtes échelles de temps évolutives à l'aide de plusieurs jeux de données. En effet, il est possible que sur de courtes échelles de temps, plusieurs gènes mobiles risquent d'être gouvernés par la dérive génétique puisqu'ils n'ont pas eu le temps d'être transférés à plusieurs populations microbiennes. Ainsi leur N_e n'est pas assez élevée pour franchir la barrière de dérive génétique et rendre leur sélection effective. Puisque ce modèle quasi neutre d'évolution peut expliquer nos résultats, il n'est pas surprenant que le taux de THG, qui influence la N_e des gènes mobiles, influence aussi largement l'évolution à court terme de ces gènes en simulation. Il serait donc intéressant de tester l'hypothèse selon laquelle la distribution de l'abondance des gènes mobiles en simulation permettrait de définir un seuil de barrière-dérive à partir duquel la sélection est effective. Selon cette hypothèse, en deçà de ce seuil, les gènes mobiles devraient accumuler plus de mutations délétères à cause d'une efficacité de sélection réduite et

dN/dS devrait être plus élevé pour ces gènes puisque la sélection purificatrice est plus faible pour ces gènes.

Enfin, la comparaison de nos résultats à la littérature suggère que l'équilibre entre les forces évolutives agissant sur le pangénome varie selon l'échelle de temps et l'unité d'évolution étudiées. De plus, nos résultats suggèrent que l'évolution des gènes mobiles serait mieux expliquée à des niveaux biologiques plus fins. En effet, on remarque qu'une plus grande partie de la variance de l'évolution des gènes mobiles est expliquée par la fonction du gène comparativement à la famille de l'hôte et encore moins par son village d'origine (**Figure 3.3 p.61**). Il serait intéressant de tester la validité de ces hypothèses.

6 Conclusion

Ce projet accorde une contribution significative à un débat en cours en microbiologie: l'évolution des pangénomes de procaryotes est-elle mieux expliquée par un modèle adaptatif ou non adaptatif? Alors que les pangénomes peuvent être largement adaptatifs à leurs hôtes bactériens sur des échelles de temps d'évolution plus longues, nous montrons que la plupart des gènes mobiles semblent se propager rapidement et égoïstement sur de courtes échelles de temps, c'est-à-dire sans nécessairement conférer un avantage sélectif immédiat à leur hôte humain. Cependant, un sous-ensemble de gènes impliqués dans des fonctions comme la défense de la cellule microbienne et son métabolisme secondaire peut être soumis à des pressions sélectives variant selon les espèces. De plus, contrairement à ce qui s'observe sur de longues échelles de temps évolutif, nous montrons que l'évolution des gènes mobiles est faiblement influencée par les attributs d'hôte, y compris ceux qui sont liés aux réseaux sociaux humains. Ce modèle n'est pas strictement neutre et laisse entrevoir qu'un modèle d'évolution quasi neutre peut expliquer nos observations, ce qui sera le sujet de futurs travaux.

Malgré le fait que notre étude se concentre sur l'évolution des pangénomes à court terme dans le microbiome intestinal humain, nous créons un cadre de modélisation simple et reproductible pour élargir l'étude des pangénomes à travers plusieurs échelles de temps et plusieurs niveaux de complexité biologique, du gène au microbiome et à l'hôte. Nous pensons donc que ce projet inspirera de futures études à étudier les facteurs qui conduisent à des changements dans l'équilibre des forces évolutives influençant le pangénome plutôt que de soutenir un modèle évolutif en particulier peu importe le contexte.

7 Références bibliographiques

Adato, O., Ninyo, N., Gophna, U., & Snir, S. (2015). Detecting Horizontal Gene Transfer between Closely Related Taxa. *PLoS Computational Biology, 11*(10), e1004408. doi:10.1371/journal.pcbi.1004408

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Andreani, N. A., Hesse, E., & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *ISME J, 11*(7), 1719-1721. doi:10.1038/ismej.2017.36

Bachtrog, D., & Gordo, I. (2004). Adaptive evolution of asexual populations under Muller's ratchet. *Evolution, 58*(7), 1403-1413. doi:10.1111/j.0014-3820.2004.tb01722.x

Bansal, M. S., Alm, E. J., & Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics, 28*(12), i283-291. doi:10.1093/bioinformatics/bts225

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Berg, J., Tymoczko, J., & Stryer, L. (2002). The Michaelis-Menten Model Accounts for the Kinetic Properties of Many Enzymes. In W. H. Freeman (Ed.), *Biochemistry 5th edition*. New York.

Bobay, L. M., & Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology, 18*(1), 153. doi:10.1186/s12862-018-1272-4

Brito, I. L., Gurry, T., Zhao, S., Huang, K., Young, S. K., Shea, T. P., . . . Alm, E. J. (2019). Transmission of human-associated microbiota along family and social networks. *Nat Microbiol, 4*(6), 964-971. doi:10.1038/s41564-019-0409-6

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., . . . Alm, E. J. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature, 535*(7612), 435-439. doi:10.1038/nature18927

Corander, J., Fraser, C., Gutmann, M. U., Arnold, B., Hanage, W. P., Bentley, S. D., . . . Croucher, N. J. (2017). Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol, 1*(12), 1950-1960. doi:10.1038/s41559-017-0337-x

Cordero, O. X., & Polz, M. F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews: Microbiology, 12*(4), 263-273. doi:10.1038/nrmicro3218

Crainiceanu, C., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66*, 165-185.

Domingo-Sananes, M. R., & McInerney, J. O. (2019). Selection-based model of prokaryote pangenomes. *biorxiv*, 782573. doi:10.1101/782573

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., & Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research, 33*(1), e6. doi:10.1093/nar/gni004

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ, 3*, e1319. doi:10.7717/peerj.1319

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews: Genetics, 8*(8), 610-618. doi:10.1038/nrg2146

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., . . . Raes, J. (2016). Population-level analysis of gut microbiome variation. *Science, 352*(6285), 560-564. doi:10.1126/science.aad3503

Gardon, H., Biderre-Petit, C., Jouan-Dufournel, I., & Bronner, G. (2020). A drift-barrier model drives the genomic landscape of a structured bacterial population. *Molecular Ecology*. doi:10.1111/mec.15628

Garud, N. R., & Pollard, K. S. (2020). Population Genetics in the Human Microbiome. *Trends in Genetics, 36*(1), 53-67. doi:10.1016/j.tig.2019.10.010

Gauthier, L., Di Franco, R., & Serohijos, A. W. R. (2019). SodaPop: a forward simulation suite for the evolutionary dynamics of asexual populations on protein fitness landscapes. *Bioinformatics, 35*(20), 4053-4062. doi:10.1093/bioinformatics/btz175

Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J, 8*(8), 1553-1565. doi:10.1038/ismej.2014.60

Hehemann, J. H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., & Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature, 464*(7290), 908-912. doi:10.1038/nature08937

Jiang, X., Hall, A. B., Xavier, R. J., & Alm, E. J. (2019). Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PloS One, 14*(12), e0223680. doi:10.1371/journal.pone.0223680

Jukes, T. H., & Cantor, C. R. (1969). Evolution of Protein Molecules. In Elsevier (Ed.), *Mammalian protein metabolism* (Vol. III, pp. 21-132). New York: Academic Press.

Koonin, E. V., & Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews: Genetics, 11*(7), 487-498. doi:10.1038/nrg2810

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2010). Detecting genomic islands using bioinformatics approaches. *Nature Reviews: Microbiology, 8*(5), 373-382. doi:10.1038/nrmicro2350

Lawrence, J. G., & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution, 44*(4), 383-397. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9089078

McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nat Microbiol, 2*, 17040. doi:10.1038/nmicrobiol.2017.40

Moulana, A., Anderson, R. E., Fortunato, C. S., & Huber, J. A. (2020). Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents. *mSystems, 5*(2). doi:10.1128/mSystems.00673-19

Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research, 28*(1), 292. doi:10.1093/nar/28.1.292

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25*(7), 1043-1055. doi:10.1101/gr.186072.114

Parsch, J., Zhang, Z., & Baines, J. F. (2009). The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. *Mol Biol Evol, 26*(3), 691-698. doi:10.1093/molbev/msn297

Ravenhall, M., Skunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Computational Biology, 11*(5), e1004095. doi:10.1371/journal.pcbi.1004095

RCoreTeam. (2019). A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sela, I., Wolf, Y. I., & Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America, 113*(41), 11399-11407. doi:10.1073/pnas.1614083113

Shapiro, B. J. (2017). The population genetics of pangenomes. *Nat Microbiol, 2*(12), 1574. doi:10.1038/s41564-017-0066-6

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research, 15*(3), 1281-1295. doi:10.1093/nar/15.3.1281

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., & Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature, 480*(7376), 241-244. doi:10.1038/nature10571

Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol, 32*(4), 1097-1108. doi:10.1093/molbev/msv003

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3), 585-595.

Takeuchi, N., Cordero, O. X., Koonin, E. V., & Kaneko, K. (2015). Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biology, 13*, 20. doi:10.1186/s12915-015-0131-7

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research, 28*(1), 33-36. doi:10.1093/nar/28.1.33

Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ, 361*, k2179. doi:10.1136/bmj.k2179

Vogan, A. A., & Higgs, P. G. (2011). The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biology Direct, 6*, 1. doi:10.1186/1745-6150-6-1

Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J., & Eyre-Walker, A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in Microbiology, 23*(10), 598-605. doi:10.1016/j.tim.2015.07.006

Wheeler, B., & M., T. (2016). Package 'lmPem': Permutation tests for linear models (Version 2.1.0). Retrieved from https://github.com/mtorchiano/lmPerm

Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E., & Koonin, E. V. (2016). Two fundamentally different classes of microbial genes. *Nat Microbiol, 2*, 16208. doi:10.1038/nmicrobiol.2016.208

Worning, P., Jensen, L. J., Nelson, K. E., Brunak, S., & Ussery, D. W. (2000). Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. *Nucleic Acids*

Research, 28(3), 706-709. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10637321

Wuitschick, J. D., & Karrer, K. M. (1999). Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. *Journal of Eukaryotic Microbiology, 46*(3), 239-247. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10377985

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., . . . Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature, 486*(7402), 222-227. doi:10.1038/nature11053

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., . . . Fu, J. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science, 352*(6285), 565-569. doi:10.1126/science.aad3369

8 Annexes

Gene	Bonferroni-adjusted p-value	COG category	Gene family	Functional description
125_15	2.09e-3	NA	PF00239	Resolvase Resolvase, N terminal domain
125_14	1.70e-2	R	COG4849	Uncharacterized protein conserved in bacteria
32371_1	2.26e-2	NA	PF13338	Protein of unknown function
18124_0	1.88e-3	P	COG4771	Outer membrane receptor for ferrienterochelin and colicins
14942_0	5.93e-3	NA	TIGR02768	Ti-type conjugative transfer relaxase TraA
25161_0	6.26e-3	H	COG0422	Thiamine biosynthesis protein ThiC
16063_1	2.65e-2	R	COG1524	Uncharacterized proteins of the AP superfamily
16796_1	4.54e-2	I	COG1022	Long-chain acyl-CoA synthetases (AMP-forming)
18960_4	7.07e-3	NA	NA	NA
23322_0	5.75e-4	S	COG3533	Uncharacterized protein conserved in bacteria

Table 3.S2A Metadata about mobile genes for which *dN/dS* significantly correlates with host household

This table presents the main functional annotations of genes for which *dN/dS* significantly correlates with host household. The Bonferroni-adjusted p-value of this correlation is also showed in the table. **NA** refers to missing data. The data were extracted from publicly available metadata from Fiji Community Microbiome Project (http://fijicomp.bme.cornell.edu//data.html).

Gene	Bonferroni-adjusted p-value	COG category	Gene family	Functional description
6633_6	4.74e-06	K	COG1191	DNA-directed RNA polymerase specialized sigma subunit
21660_2	3.59e-03	NA	NA	Protein of unknown function
20356_1	4.42e-06	M	COG0744	Membrane carboxypeptidase (penicillin-binding protein)
28857_261	3.22e-03	X	COG2801	Transposase and inactivated derivatives
23847_1	5.36e-06	NA	PF09839	DUF2066 Uncharacterized protein conserved in bacteria (DUF2066)
32729_2	2.25e-04	G	COG0366	Glycosidases
125_19	1.03e-05	NA	NA	NA
16377_13	7.18e-04	NA	NA	Protein of unknown function
18097_1	3.01e-03	R	COG2819	Predicted hydrolase of the alpha/beta superfamily
14160_7	4.11e-03	V	COG1131	ABC-type multidrug transport system, ATPase component
74_6	7.10e-06	NA	PF01638	HTH HxlR HxlR-like helix-turn-helix
445_1	4.48e-02	NA	PF13529	Peptidase_CA Peptidase_C39_2 Peptidase_C39 like family
3196_18	1.80e-03	NA	PF12844	HTH_19 Helix-turn-helix domain
3196_27	4.63e-04	NA	PF14283	Protein of unknown function
6633_7	3.98e-02	K	COG1595	RNA polymerase sigma factor sigma-70 family
8611_88	1.89e-02	NA	PF13149	Protein of unknown function
14942_2	1.29e-03	S	COG5658	Predicted integral membrane protein
17873_1	1.61e-05	M	COG0472	UDP-N-acetylmuramyl pentapeptide phosphotransferase/ UDP-N- acetylglucosamine-1-phosphate transferase

17842_0	3.24e-02	F	COG0150	Phosphoribosylaminoimidazole (AIR) synthetase
26824_154	9.56e-06	Q	COG0656	Aldo/keto reductases, related to diketogulonate reductase
28857_171	1.87e-02	E	COG0687	Spermidine/putrescine-binding periplasmic protein
28857_290	2.44e-03	E	COG0169	Shikimate 5-dehydrogenase
28689_7	2.51e-02	H	COG0379	Quinolinate synthase
26824_56	3.91e-03	NA	NA	Protein of unknown function
33358_13	1.39e-02	NA	NA	Protein of unknown function
25426_7	4.14e-02	H	COG0422	Thiamine biosynthesis protein ThiC
125_6	8.58e-09	L	COG4974	Site-specific recombinase XerD
26824_98	3.22e-02	C	COG0045	Succinyl-CoA synthetase, beta subunit
24698_0	1.38e-02	E	COG0128	5-enolpyruvylshikimate-3-phosphate synthase
27262_1	4.45e-02	O	COG0545	FKBP-type peptidyl-prolyl cis-trans isomerases 1
24455_0	4.68e-02	K	COG2865	Predicted transcriptional regulator containing an HTH domain and an uncharacterized domain shared with the mammalian protein Schlafen
23452_0	1.61e-06	NA	K12071	conjugal transfer protein TraD
24325_0	2.99e-02	V	COG0534	Na+-driven multidrug efflux pump

Table 3.S2B Metadata about mobile genes for which *Tajima's D* significantly correlates with host household

This table presents the main functional annotations of genes for which *Tajima's D* significantly correlates with host household. The Bonferroni-adjusted p-value of this correlation is also showed in the table. **NA** refers to missing data. The data were extracted from publicly available metadata from Fiji Community Microbiome Project (http://fijicomp.bme.cornell.edu//data.html).

Gene	Bonferroni-adjusted p-value	COG category	Gene family	Functional description
28857_123	4.96e-02	J	COG2265	SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase
20356_11	3.60e-04	G	COG0057	Glyceraldehyde-3-phosphate dehydrogenase/ erythrose-4-phosphate dehydrogenase
35737_1	3.10e-02	C	COG1053	Succinate dehydrogenase/ fumarate reductase, flavoprotein subunit
35737_15	8.22e-03	C	COG0479	Succinate dehydrogenase/ fumarate reductase, Fe-S protein subunit
28857_39	2.27e-03	NA	TIGR01391	DNA replication, recombination and repair DNA primase, catalytic core
35631_13	8.06e-03	IQ	COG0236	Acyl carrier protein
28857_299	4.03e-04	QR	COG0500	Protein of unknown function
25368_7	6.89e-05	P	COG0581	ABC-type phosphate transport system, permease component
28857_57	2.05e-02	G	COG0366	Glycosidases
29922_2	2.82e-02	G	COG0176	Transaldolase
24455_2	4.83e-02	M	COG4591	ABC-type transport system, involved in lipoprotein release, permease component
28857_267	8.66e-05	E	COG0685	5,10-methylenetetrahydrofolate reductase
25426_5	1.58e-04	IQ	COG0318	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II

32159_6	4.73e-04	H	COG1541	Coenzyme F390 synthetase
26824_28	3.71e-06	T	COG2905	Predicted signal-transduction protein containing cAMP-binding and CBS domains

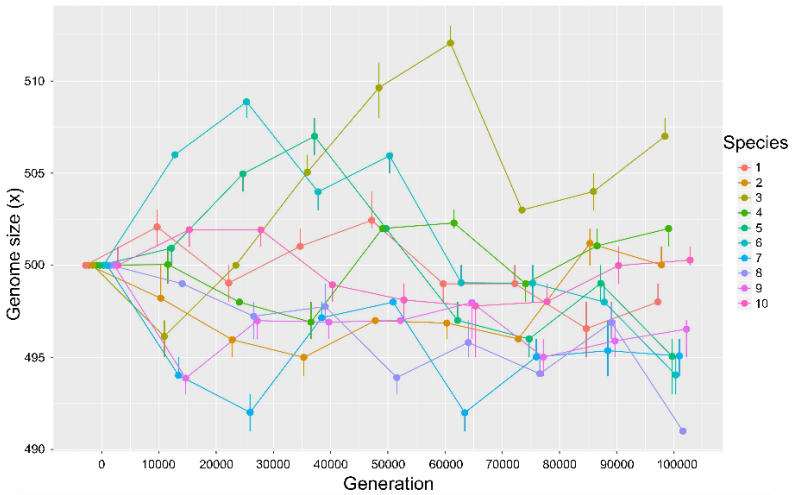
Table 3.S2C Metadata about mobile genes for which θ_π significantly correlates with host Village

This table presents the main functional annotations of genes for which θ_π significantly correlates with host Village. The Bonferroni-adjusted p-value of this correlation is also showed in the table. **NA** refers to missing data. COG categories that are related to the subset of functions that Brito et al. (2016) identified as being particularly abundant or prevalent in certain Fiji villages are highlighted in yellow. The data were extracted from publicly available metadata from Fiji Community Microbiome Project (http://fijicomp.bme.cornell.edu//data.html).

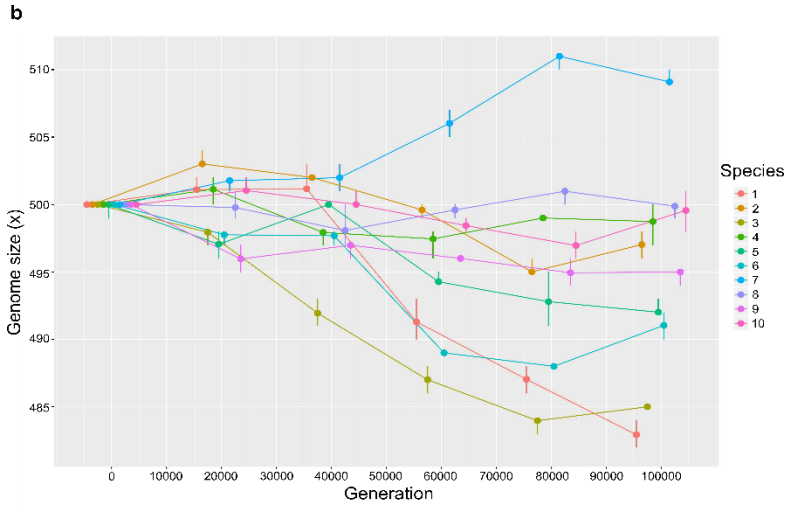
Figures 3.S10 Genome size equilibrium across simulations

This set of figures represents the time series of genome size during the simulations mentioned in the manuscript presented in section 3. The simulations reach a genome size equilibrium around 500 genes. It is important to show that these time series are on dynamic equilibrium, i.e. genome size fluctuates around a certain value, because it supports the fact that our results are not dependent on the initial conditions of the simulations:

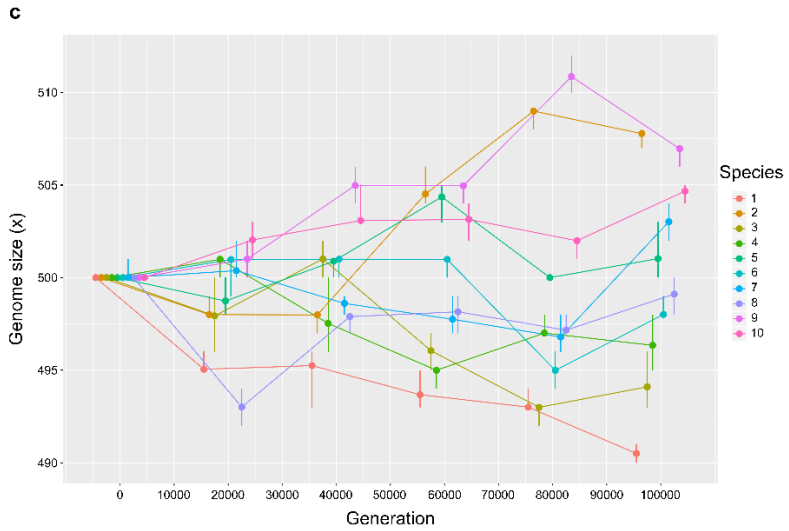
a



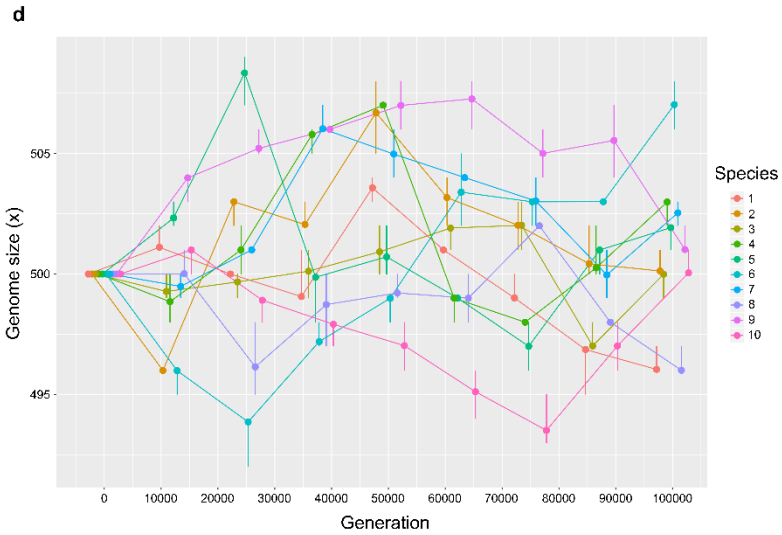
a) HGT rate = 0.01μ and HGT is neutral.



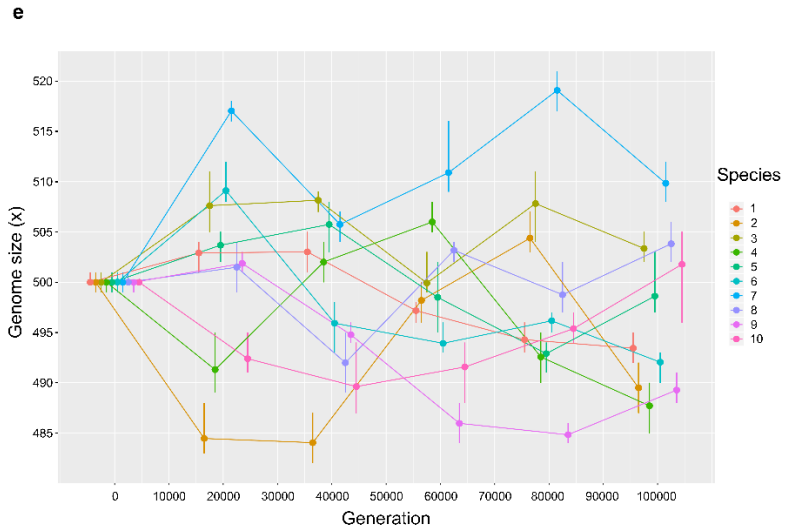
b) HGT rate = 0.01μ and λ = 1E6



c) HGT rate = 0.01μ and λ = 1E5

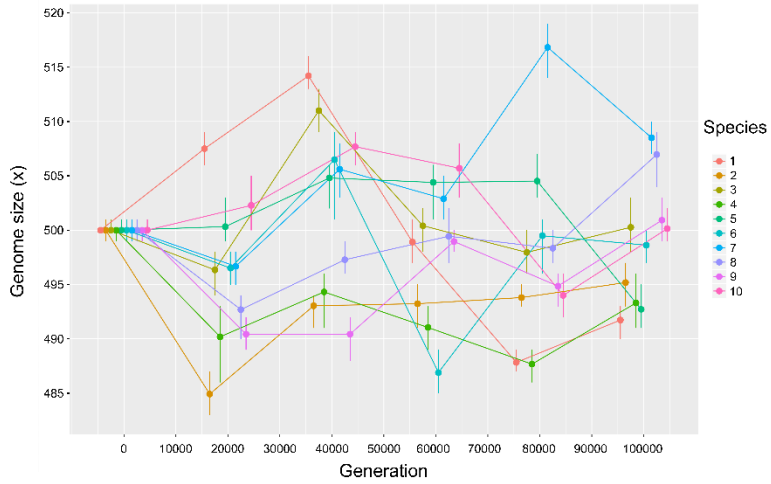


d) HGT rate = 0.01μ and λ = 1E4



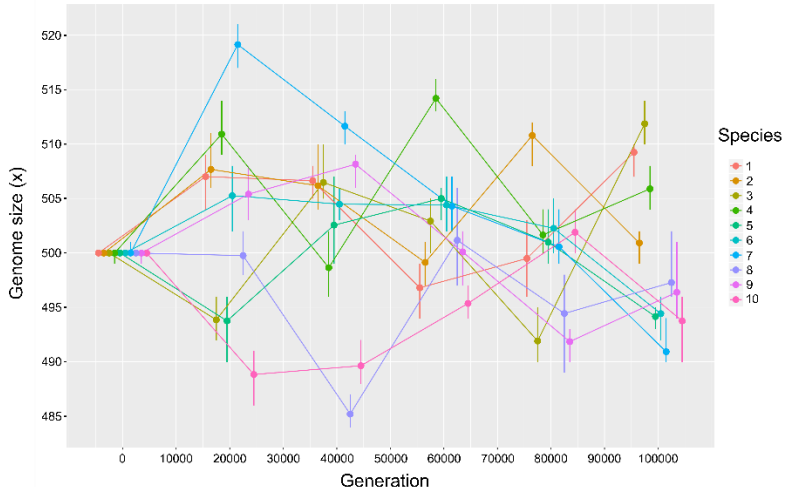
e) HGT rate = 0.1μ and HGT is neutral

f

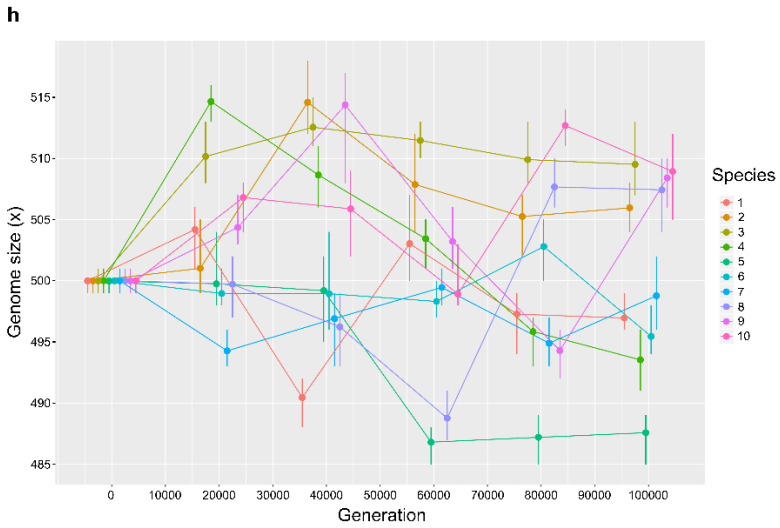


f) HGT rate = 0.1μ and λ = 1E6

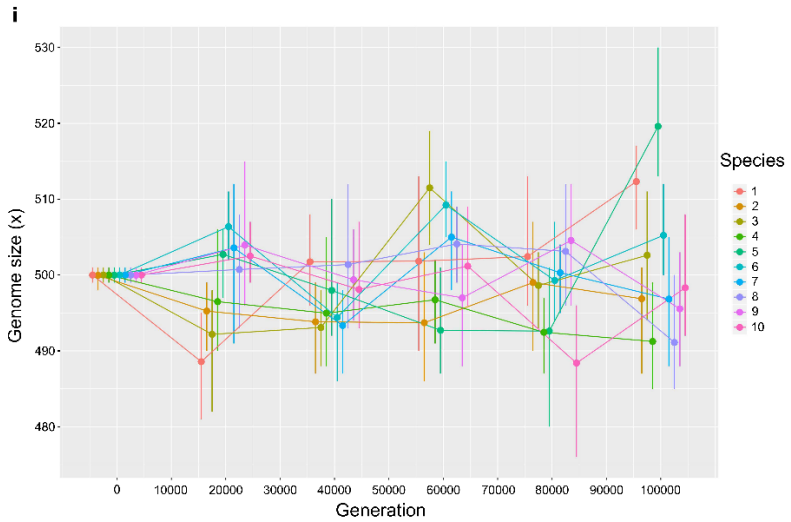
g



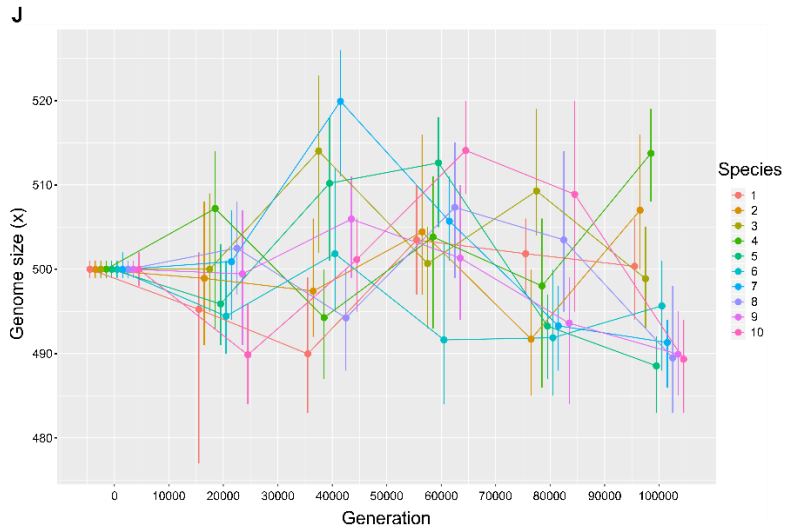
g) HGT rate = 0.1μ and λ = 1E5



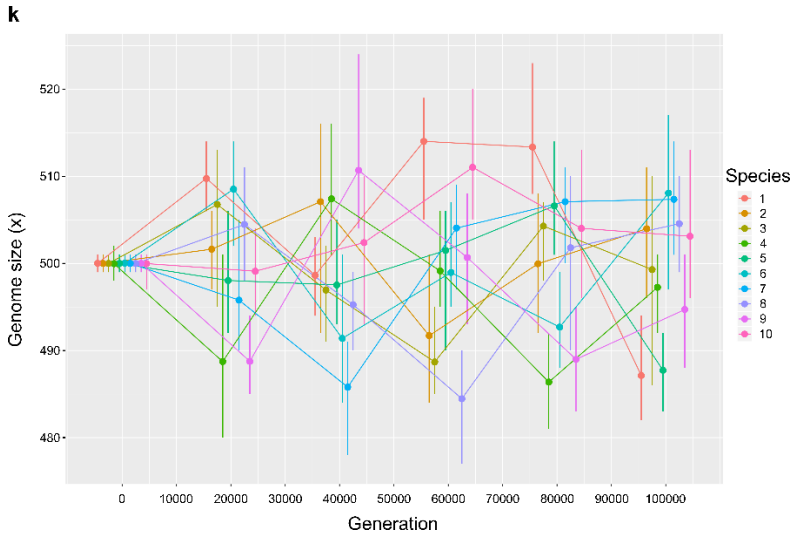
h) HGT rate = 0.1μ and λ = 1E4



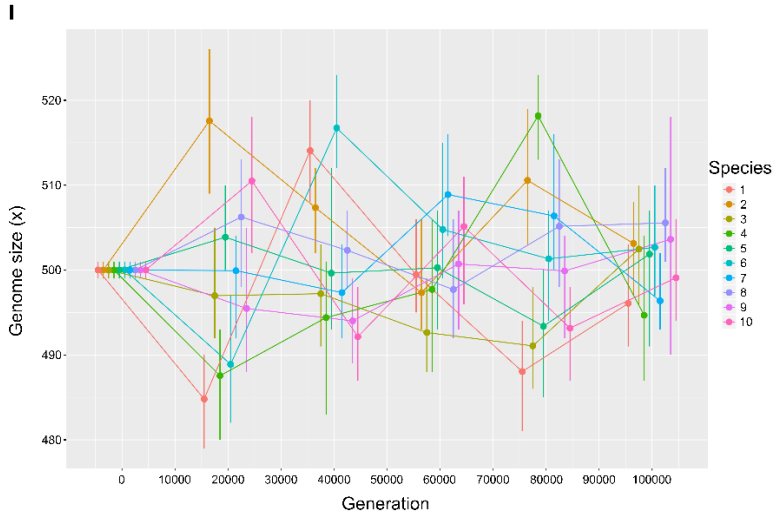
i) HGT rate = 1μ and HGT is neutral



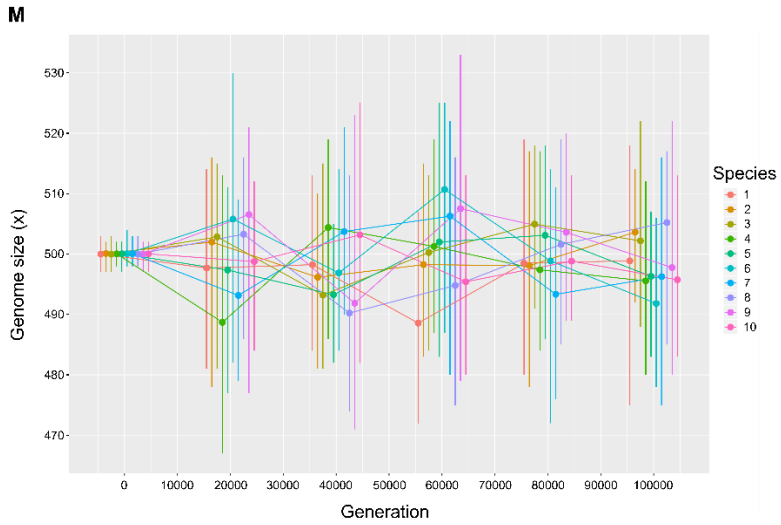
j) HGT rate = 1μ and λ = 1E6



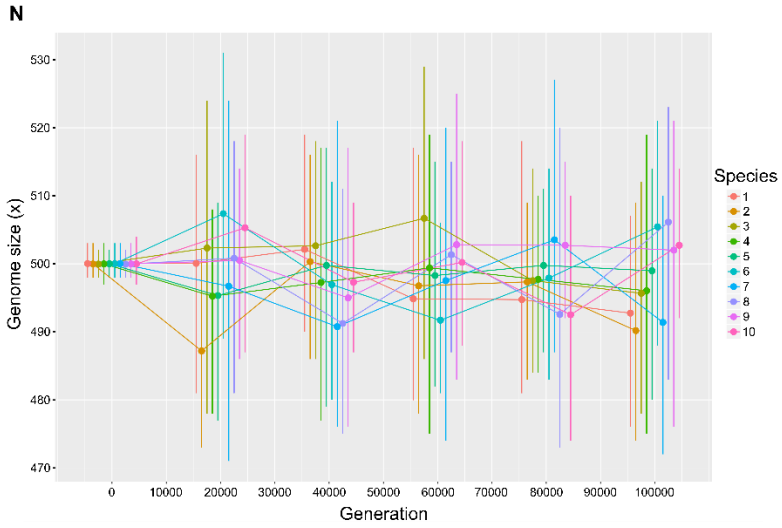
k) HGT rate = 1μ and λ = 1E5



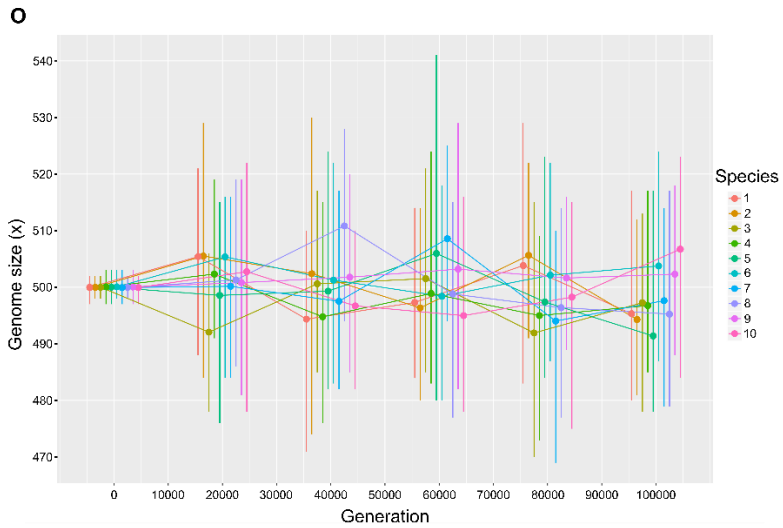
l) HGT rate = 0.01μ and λ = 1E4



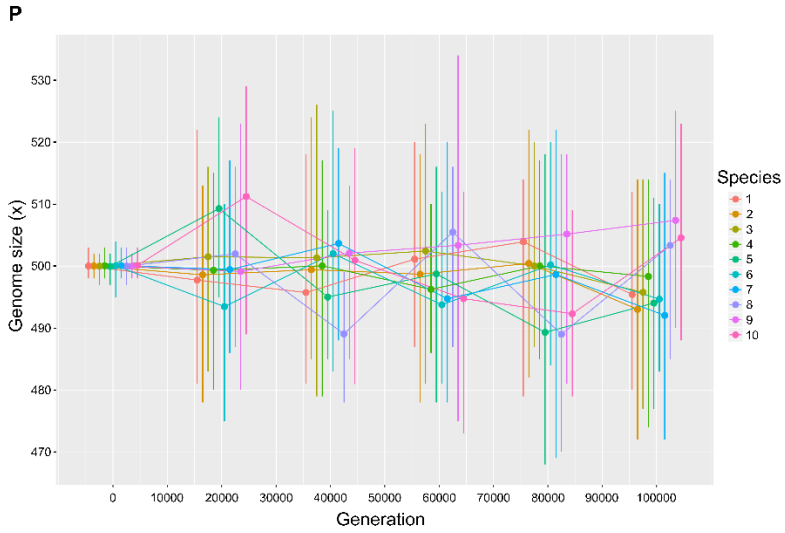
m) HGT rate = 10μ and HGT is neutral



n) HGT rate = 10μ and λ = 1E6



o) HGT rate = 10μ and λ = 1E5



p) HGT rate = 10μ and λ = 1E4